Rapidly evolving pathogens in a polymorphic host population

Modeling immuno-epidemiology of HIV-1 and influenza A virus

Christiaan H. van Dorp

August 31, 2018

Cover:

detail: *New Skin for the Old Ceremony v. III* (2018)
Doreen Wittenbols
water-soluble oil on paper

ISBN: 978-94-9301-426-8
Printed by Gildeprint, Enschede

Rapidly evolving pathogens in a polymorphic host population

Modeling immuno-epidemiology of HIV-1 and influenza A virus

Snel evoluerende pathogenen in een polymorfe gastheerpopulatie

Modellen van de immuno-epidemiologie van griep en AIDS

(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de Universiteit Utrecht op gezag van de rector magnificus, prof.dr. H.R.B.M. Kummeling, ingevolge het besluit van het college voor promoties in het openbaar te verdedigen op vrijdag 31 augustus 2018 des middags te 4.15 uur

door

Christiaan Hendrik van Dorp

geboren op 11 januari 1986 te Haarlem

Promotor: Prof.dr. R.J. de Boer

Copromotor: Dr. M. van Boven

Contents

Chapter 1

Introduction

Pathogens can adapt to their hosts on multiple levels. At the between-host level, natural selection acts on infectiousness, avoiding immunological memory in the host population, and pathogen-induced mortality. At the within-host level, pathogens are selected for immune escape and pathogen load. These selection pressures do not necessarily operate in concert, and this is further complicated by the massive host-heterogeneity that probably evolved as a defense to pathogens. In humans, the most striking example of heterogeneity is found in the major histocompatibility complex (MHC) molecules, which are cell-surface proteins used by the immune system to present epitopes to T cells. In this thesis, we study the immunology and epidemiology of two pathogens, one causing acute infections (influenza A virus; IAV), the other inducing a chronic infection (human immunodeficiency virus type 1; HIV-1). In both cases, we are interested in how these pathogens evolve in a heterogeneous host population. In particular, we study the escape from immune responses that are restricted by the highly polymorphic MHC molecules, and the consequences of escapes on the population level.

1.1 Host heterogeneity and HLA polymorphism

MHC molecules are key elements of the adaptive immune system. MHC class I molecules present intracellular peptides to $CD8^+$ T cells, provoking an immune response when the T-cell receptor (TCR) binds the peptide-MHC (pMHC) complex. Intracellular peptides are derived from mis-folded proteins and possibly originate from viruses if the cell is infected. TCRs are generated at random using a process of recombination and nucleotide insertions, and T-cells with TCRs that bind any self-peptide MHC complex are negatively selected during their development in the thymus.

Human leukocyte antigens (HLA), the human MHC molecules, are incredibly polymorphic. At the time of writing, 13 324 HLA class I alleles have been documented

[www.ebi.ac.uk/ipd/imgt/hla; 190]. This number should be taken with a grain of salt, as most of these alleles are very rare, and many alleles can functionally be similar [189]. HLA molecules only bind a small subset of all possible peptides, and the makeup of this subset differs from one molecule to the next.

The evolutionary origins of the massive HLA polymorphism is under active debate, as there are multiple potential mechanisms that could generate this diversity, and it is difficult to pinpoint the mechanisms responsible for balancing selection using genomic data [150]. One candidate is a heterozygote advantage. Heterozygous individuals are able to present a wider range of epitopes. Therefore they should be able to mount more immune responses against a pathogen, giving them a fitness benefit. This mechanism can maintain multiple HLA alleles in a population, but only to a limited extent. Another candidate mechanism is negative frequency-dependent selection. Since a pathogen is more likely to escape from responses restricted to common HLA alleles, individuals with these common alleles will get a fitness disadvantage, while individuals with rare alleles that the pathogen has not escaped from will gain a relative fitness advantage.

1.2 Influenza A virus

The first part of this thesis dives into the evolution of $CD8^+$ T-cell epitopes of influenza A virus (IAV). IAV is a single stranded RNA virus with a segmented genome. The 8 segments of IAV have a total length of 13 588 bases and code for 10 to 14 proteins. Two surface proteins, hemagglutinin (HA) and neuraminidase (NA) are used to designate the many subtypes of IAV. Humans are predominantly infected by the subtypes H1N1 and H3N2, which cause yearly seasonal epidemics in the temperate regions. IAV is able to persist in the human population, in spite of the fact that it can induce sterilizing immunity, because of antigenic drift and shift. Antigenic drift is discussed in detail below. Antigenic shift can occur when a host is infected with multiple IAV subtypes, and gene segments are re-assorted. After the introduction of the H1N1 subtype in 1918 that caused a pandemic killing 50 to 100 million people, re-assortment has led to (less severe) pandemics in 1957 (H2N2), 1968 (H3N2), and 2009 (H1N1) [156].

1.2.1 Antigenic cartography of the influenza virus

In 2004, an influential paper was published by Smith *et al.* [210], that introduced a broad audience to antigenic cartography of the influenza A virus. The idea of antigenic cartography of influenza was introduced a couple of years before that [119], but Smith *et al.* [210] were able to make use of a much larger dataset, resulting in a very clear picture (Figure 1.1A). In short, in antigenic cartography influenza strains and immune responses are represented by points in a low dimensional space (typically the

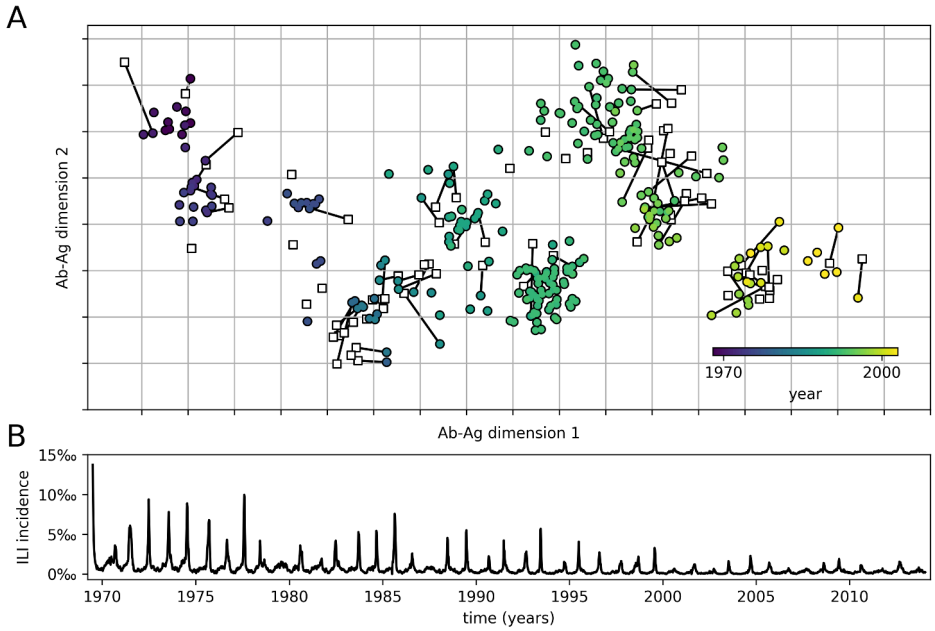


Figure 1.1: Antibody-antigenic map of IAV (H3N2) and the incidence of ILI. A. The coordinates of immune responses (squares) and influenza strains (circles) were calculated by Smith *et al.* [210], and downloaded from `www.antigenic-cartography.org`. An immune response and the strain that was used to induce the immune response are connected by a line. The distance between grid lines corresponds to a 4 fold increases of the titer. **B.** Time series of influenza-like illness in the Netherlands. The weekly incidence equals the number of ILI cases reported by GPs, divided by the sentinel population size.

dimension is 2), and the distances between immune responses and influenza strains are representative of the strength of these responses against the strains.

The data needed to create antigenic maps is based upon hemagglutination inhibition (HI) assays. Influenza virus uses sialic-acid receptors to bind and enter the host cell. This binding is mediated by the HA envelope protein. Sialic acid can be found on the surface of many cells, including erythrocytes. A large enough concentration of influenza virus causes erythrocytes to form a lattice, a process called hemagglutination, which can easily be measured by visual cues. Typically, a ferret is infected with an influenza strain v_i, the ferret mounts an immune response r_i, and antibodies against the virus are isolated from the ferret. IgA antibodies that target HA can cause sterilizing immunity by inhibiting the virions from entering the target cells. This process is simulated in the HI assay: If the ferrets' antibodies are specific for the HA of the virus, the hemagglutination is inhibited by adding antibodies to the erythrocytes. This will typically happen for the

strain v_i that was used to infect the ferret in order to produce the antibodies r_i, but the usefulness of the HI assay becomes clear when the HI of a different antibody r_j is measured for strain v_i. This process results in a matrix (H_{ij}) that represents the cross-reactivity of antibody responses for different influenza strains. The quantity H_{ij} is the maximum dilution of the antiserum r_i at which hemagglutination is inhibited, i.e. the titer.

The key to antigenic cartography is interpreting the titer as a distance in an abstract antigenic space (or shape space), a concept introduced by Perelson and Oster [174]. More precisely, the used distance is

$$d_{ij} = \log_2(H_{\max,j}) - \log_2(H_{ij}), \tag{1.1}$$

where the base 2 logarithm is used because the dilutions are two-fold. The maximum titer $H_{\max,j} := \max_k(H_{kj})$ makes sure that the distances are non-negative. Notice that the distances d_{ij} are not necessarily symmetric (i.e. $d_{ij} \neq d_{ji}$). In order to convert a distance matrix (d_{ij}) into a map, a method to reduce the dimensionality is used, typically multidimensional scaling (MDS). For MDS, every data point (v_i for viruses and r_j for immune responses) is represented by a point in Euclidean space (x_i and y_j respectively), and one defines an error function

$$E(x,y) = \sum_{i,j} \left(d_{ij} - \|x_i - y_j\|_2\right)^2, \tag{1.2}$$

where $\|x_i - y_j\|_2$ denotes the Euclidean distance between the points x_i and y_j. Different error functions correspond to different flavors of MDS; the one discussed here is termed metric MDS. In order to find the best representatives x_i and y_j, the error function has to be minimized, typically by some gradient descent method.

MDS is not merely a data visualization method; it is also used to infer missing distances. The HI data matrix used for by Smith *et al.* [210] is quite sparse as many of the titers between antigen v_i and antiserum r_j are not determined. When the antigenic distance between antigen and antiserum is too large, the HI assay is not sensitive enough to produce a meaningful value. Moreover, some of the titers that are determined are left or right censored (in fact, all HI titer data is interval censored by design of the experiment). Furthermore, the antigenic map results in distances between antigens (and also distances between antisera) which are not present in the original data.

The antigenic map of IAV has a couple of interesting characteristics. First, the locations of the strains appear to have a certain direction. With increasing time (indicated by hue in Figure 1.1A), the strains follow more or less a straight line. Although there has been some discussion about whether this is an artifact of MDS [155], more often this pattern is attributed to antigenic drift of the virus. After each IAV epidemic, many individuals have become immune to subsequent infection with the same IAV strain. The

idea is that the IAV mutant that is antigenically the farthest away from previous strains, has the best chance of seeding the next epidemic. This idea may also underlie the surprising lack of branching in the phylogenetic tree of IAV, although the development of a parsimonious model that mimics this behavior has taken considerable effort [111].

Secondly, the optimal dimension of the target space for MDS does not need to be 2, and it might be that a higher dimensional target space results in a much better fit, as evaluated by cross-validation. For the antigenic map of influenza, it appears that increasing the dimension only marginally improves the fit. Hence, a 2-dimensional Euclidean space turns out to be a good representation of the antigenic space of influenza, which has been used to model the genetic evolution of IAV [11]. Finally, the map clearly shows antigenic clusters of the isolates. This pattern has been described as punctuated evolution, and is contrasted by the gradual genetic evolution of the virus [210].

Recently, more advanced versions of MDS have been developed that can use additional data to inform the antigenic map. Bayesian MDS (BMDS), as developed by Bedford *et al.* [12], allows for the integration of genetic data to inform the location of viruses on the map. BMDS differs from MDS in the sense that the error function is replaced with a likelihood function, and the locations of viruses on the map are inferred with a Markov chain Monte Carlo (MCMC) approach. An additional benefit of the likelihood approach is that it is well suited for handling the intrinsically censored HI data. The evolution of the antigenic phenotype of a strain is modeled as a diffusion process on the phylogenetic tree of the virus. Hence, the diffusion process serves as a prior distribution for the location of the strains. Alternatively, the sampling date is used as prior information. In this case, antigenic drift is added explicitly to the model, and the prior distribution of the location x_i of strain v_i that has been isolated at time t_i is given by

$$x_i \sim \mu \cdot \binom{t_i}{0} + \mathcal{N}_2(0, \sigma_v^2 I_2)\,. \tag{1.3}$$

Here, I_2 is the 2×2 identity matrix. The rate of the antigenic drift is given by the parameter μ, and the volatility by σ_v. By design, antigenic drift occurs predominantly in the direction of the first antigenic dimension. In Chapter 3 we refer to this prior distribution as the drift prior.

In another interesting MDS variant, the two-dimensional space that represents antigenic space is replaced by a space that is directly determined by the phylogenetic tree [166]. In this model, each branch of the tree is equipped with a parameter representing the antigenic distance between parent node and child node. Each strain is represented by a leaf of the tree, and the distance between two strains is given by the sum of the antigenic distances on the shortest path between the representative leafs. The antigenic distances on the branches are fitted to the HI data with a nonlinear optimization algorithm. These advanced antigenic cartography methods have been used to track and predict antigenic evolution of the influenza virus in real-time [`nextstrain.org`; 165], and serve as

a basis for the World Health Organization to make decisions about vaccine updates [157], which are essential for an effective vaccine against seasonal influenza.

1.2.2 T-cell antigenic drift

Antigenic cartography has not only been applied to antibody responses against influenza virus, but also to T-cell responses [24]. The HLA-B*35 restricted IAV T-cell epitope $NP_{418\text{-}426}$ naturally occurs in a relatively large number of variants (see Chapter 2), each potentially inducing a distinct T-cell response. Boon *et al.* [24] generated a small number of CTL clones using blood from HLA-B*35$^+$ donors. The functional avidity of each of these CTL clones r_j was then measured using target cells pulsed with each variant v_i of the $NP_{418\text{-}426}$ epitope at different concentrations. The peptide concentration at which 50% of the target cells are killed (the EC_{50} value) was used as an avidity measure A_{ij}. The avidity can be transformed into a distance measure $d_{ij} = c + \log_{10}(A_{ij})$, for some constant c to ensure that all distances are non-negative. Using these distances, Boon *et al.* [24] use MDS to construct a T-cell antigenic map of IAV for the $NP_{418\text{-}426}$ epitope, i.e., the map shows the antigenic dissimilarity among the different variants of this epitope.

Each point of the T-cell antigenic map constructed by Boon *et al.* [24] represents a number of different IAV strains (that share the variant of the epitope). Hence, the map only represents the dissimilarities of these strains at one T-cell epitope. In comparison, the distances in the antibody-antigenic map from Smith *et al.* [210] represent dissimilarities based on 5 different epitopes on HA, and these epitopes are much more variable than the $NP_{418\text{-}426}$ epitope. A similarly complete T-cell antigenic map of IAV would have an advantage over the antibody-antigenic map, as different subtypes of IAV can be compared using T-cell responses, while this is not possible for antibody responses. T-cell responses can be cross-reactive between subtypes, and hence a prior infection with a H3N2 virus can lead to T-cell memory against a pandemic H1N1 strain [213].

In Chapter 2, we create a new T-cell antigenic map that takes multiple CD8$^+$ T-cell epitopes into account. We start by compiling a list of all experimentally observed epitopes. While the $NP_{418\text{-}426}$ epitope is HLA-B*35 restricted, our list contains epitopes that are restricted to many more HLA alleles, together covering a large part of the population. We use the protein sequences of a large number of historical IAV isolates to determine which epitope is present in what virus, and convert this binary matrix into a distance matrix between viruses. This matrix can be used for MDS.

The sterilizing immunity mediated by IgA antibodies against HA leads to strong positive selection on HA, and IAV is capable of causing annual epidemics because of antibody-antigenic drift. Whether similar evolutionary dynamics exist for T-cell epitopes is

not as clear. It turns out that it is difficult to detect positive selection within T-cell epitopes, because HLA molecules preferentially bind more conserved peptides of the influenza proteome [92]. Notice the contrast with the epitopes on the HA glycoprotein, which is antigenically highly evolvable, as facilitated by its "inherent tolerance for mutations" [221]. Fortunately, nature has provided a null model with swine influenza virus, that shares a common ancestor (that existed around the year 1918) with human influenza virus for the matrix protein (M1 and M2) and nucleoprotein (NP) segments. By comparing substitution rates of human to swine viruses, within and outside HLA-restricted $CD8^+$ T-cell epitopes, Machkovech *et al.* [140] were able to provide evidence for positive selection on these epitopes.

In Chapter 2 we find a similar result by looking at the number of epitopes that are present in IAV protein sequences over time. We find that the H3N2 lineage on average loses one $CD8^+$ T-cell epitope every 3 years, which can be interpreted as slow escape from human $CD8^+$ T-cell responses. We then introduce a measure of T-cell antigenic drift in terms of the number of epitopes that disappears each year.

1.2.3 Feedback on the epidemiological dynamics

Antigenic drift in both antibody and T-cell antigenic space is "punctuated". From one epidemic to the next, the jump in antigenic space is large in some years, and small in others. A natural hypothesis is that the size of these epidemics co-varies with the size of the antigenic jump. Interestingly, evidence for this natural relation is scarce. Only Bedford *et al.* [12] were able to find a significant correlation between the attack rate and the size of the jump in antibody-antigenic space.

In the Netherlands, a sentinel network of general practitioners (GPs) has been reporting weekly cases of influenza-like illness (ILI) since the year 1970. From the resulting time series (Figure 1.1B), it is possible to estimate epidemiological parameters such as the basic reproduction number and the fraction of the population that is susceptible at the start of each seasonal epidemic. Previously, the Dutch ILI time series has been used to identify the driving factors of influenza transmission [13]. A large part (27%) of the variation of the transmission intensity (the reproduction number) was explained by unknown between-season effects, which motivated us to investigate if at least part of this 27% could be explained by antigenic drift, and in particular T-cell antigenic drift.

In Chapter 3 we use MCMC to fit a model of Kermack-McKendrick type to the ILI time series. The model describes the size of susceptible (S), infected (I), and recovered (R) compartments in the Dutch population, each stratified into 6 age classes. Social contact data between individuals of different ages is used to inform the rate of infectious contact between the age classes [226]. The model parameter of interest is the fraction of susceptible individuals ($S_0 = S(t_0)$) at the start (t_0) of 45 seasonal influenza epidemics

(between 1969 and 2014). The estimated fraction of susceptible individuals is highly variable between years and age classes, but we find no evidence that this variation is due to antibody- or T-cell antigenic drift of the virus.

1.3 HIV-1

In the second part of this thesis, we focus on an example of a chronic virus: HIV-1. HIV-1 is a retrovirus with a genome of 9719 bases that encodes 16 proteins. HIV-1 uses the CD4 molecule as a receptor to bind and enter its target cells, and primarily infects $CD4^+$ T-cells and macrophages, depending on the tropism of the virus. An untreated HIV-1 infection can roughly be separated into three phases, the acute, chronic (or asymptomatic), and AIDS phase. During the acute phase, the virus load (i.e., the number of virions per ml of blood) reaches a peak within the first few weeks, after which it decreases to a lower value, the set-point virus load (SPVL). During the chronic phase, the virus load remains relatively stable. However, the $CD4^+$ T-cell count slowly decreases, eventually leading to immune failure and opportunistic infections, i.e., acquired immune deficiency syndrome (AIDS). During the AIDS phase, the virus load increases again. The stable chronic-phase virus load within patients falls into stark contrast with the large variation of the SPVL between patients. The SPVL can differ 4-5 orders of magnitude between patients (which is comprehensibly visualized by Fraser *et al.* [66]).

1.3.1 Evolution of the set-point virus load of HIV-1

Our interest in HIV-1 immuno-epidemiology was sparked by a paper published in 2007, in which Fraser *et al.* [65] made the argument that the SPVL of HIV-1 had evolved to values that maximize the basic reproduction number ($\mathscr{R}_0$). It was known from earlier studies that patients with a high SPVL tend to be more infectious than patients with low SPVLs [61]. At the same time, patients with a higher SPVL progress to AIDS more rapidly than patients with low SPVLs. Hence, the duration of the infection is shorter for patients with a high SPVL than those with low SPVLs, but per unit of time, the patients with the high SPVL transmit more.

The relationship between SPVL, transmission rate, and the duration of the asymptomatic phase is described by Fraser *et al.* [65] in a more rigorous manner than formulated above. The link between SPVL and infectiousness is given by an increasing Hill function $\beta_\theta(V)$ (see Figure 1.2A), which is fitted to data on serodiscordant couples from Zambia [61]. The relation between the SPVL and the length of the asymptomatic phase is modeled with a decreasing Hill function $D_\theta(V)$ (see Figure 1.2B), which in turn is fitted to data from seroconverters from the Netherlands [237]. The product of the

infectiousness $\beta_\theta(V)$ and the duration $D_\theta(V)$ gives the expected number number of secondary infections caused by an individual with SPVL V during the asymptomatic phase. This product—referred to as the transmission potential (TP)—is closely related to $\mathscr{R}_0$ (but not identical).

In Figure 1.2C we have plotted the transmission potential together with a histogram of a sample of SPVLs. The sample is simulated, but closely resembles SPVLs found in a real cohort [237]. The TP has a single optimum at intermediate SPVLs, and approaches 0 for small and large SPVL. The biological interpretation is that although patients with a low SPVL are infectious for a very long time, their infectiousness per unit of time is too small to transmit the virus to a meaningful number of new hosts. On the other hand, patients with a high SPVL are more infectious, but they do not live long enough to transmit the virus. Patients with intermediate SPVLs have the ideal combination (from the perspective of the virus) of sufficient infectiousness and a long enough asymptomatic phase to cause many new infections. Strikingly, patients with such an intermediate SPVL are the most common, as the peak of the SPVL distribution coincides with the peak of the TP.

Fraser *et al.* [65] argued that this is not a coincidence, but a result of natural selection: HIV-1 has evolved to attain average SPVLs that optimize the number of secondary infections. It was assumed that the variation in $\log_{10}$ SPVL is the result of host, environmental or random effects (e), but that the mean $\log_{10}$ SPVL (g) is under genetic control of the virus. If we ignore transmission during the acute and AIDS phase, the $\mathscr{R}_0$ of this virus can then be calculated by averaging over the environmental factor e:

$$\mathscr{R}_0(g) = \int_{-\infty}^{\infty} \varphi(e+g)\beta(e+g)D(e+g)\mathrm{d}e\,, \tag{1.4}$$

where φ is the probability density function of the SPVL skew-normal distribution. The observation made by Fraser *et al.* [65] is that the trait g that maximizes $\mathscr{R}_0$, is very close to the average SPVL observed in their cohort.

The evolution of the SPVL towards an $\mathscr{R}_0$-optimizing value can be understood in the framework of adaptive dynamics. In adaptive dynamics, the simplifying assumption is made that the timescales of the evolution of a trait or strategy and the ecological dynamics are separated; a new mutant only appears in the population after the previous mutant has gone to fixation. The evolutionary dynamics can be analyzed using a pairwise invasibility plot (PIP; Figure 1.2D). In order to illustrate this, we consider a very simple HIV-1 population-level model. Let S denote the number of susceptible individuals and I the number of infected patients. Susceptible individuals are born at a constant rate s, and die at a *per-capita* rate μ. Infected individuals die at a higher rate $\mu + \delta(V)$, where $\delta = D^{-1}$, and susceptible individuals are infected with a V-virus at rate $\beta(V)I/N$. Here $N = s/\mu$ is the size of the "virgin" (i.e. uninfected) population.

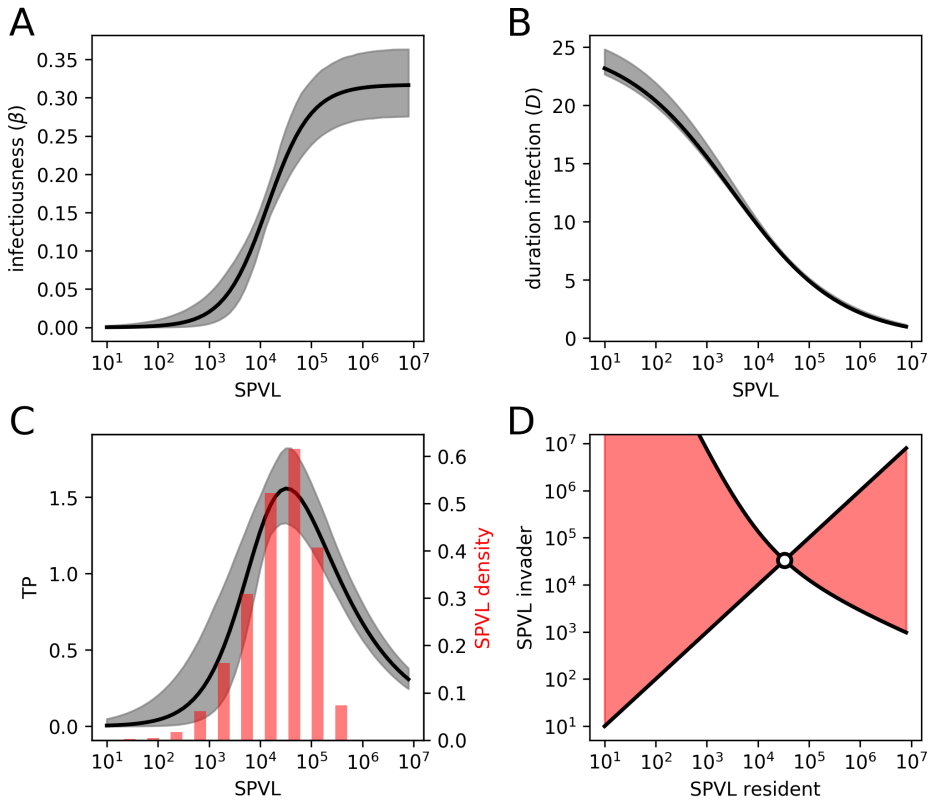


Figure 1.2: Life-history trade-off of HIV-1. A. An increasing Hill function describes the relation between SPVL and infectiousness. **B.** A decreasing Hill function describes the relation between SPVL and the length of the asymptomatic phase. **C.** The transmission potential (TP) during the asymptomatic phase as a function of SPVL. The red histogram represents a simulated SPVL sample coming from a skew-normal distribution with parameters estimated by Fraser *et al.* [65]. **D.** A PIP based on the life-history trade-off of HIV-1. The red areas indicate that the mutant can invade. The dot indicates the ESS. The gray bands depict the 95% confidence intervals (CIs) at each SPVL value. The CIs are estimated by simulation. First, we re-sample a set of $\log_{10}$ SPVLs from the afore-mentioned skew-normal distribution of size 1000. Then, we simulate for each SPVL a duration of the asymptomatic phase, and a Boolean deviate indicating whether a transmission occured within a time interval of one year. The durations and Boolean deviates are then used to re-estimate the parameters (θ) for the asymptomatic duration function $D(V)$ and the infectivity function $\beta(V)$, respectively. This results in an ensemble of 1000 parameter sets θ. At each SPVL V, we compute $D_\theta(V)$ and $\beta_\theta(V)$, from which we derive the 2.5 and 97.5 percentiles.

We assume that the virus and host population is monomorphic, and that V is every patient's SPVL. The model is given by the following ordinary differential equations (ODEs) in which t denotes time:

$$\begin{aligned}\frac{\mathrm{d}S}{\mathrm{d}t} &= s - \mu S - \beta(V)SI/N \\ \frac{\mathrm{d}I}{\mathrm{d}t} &= \beta(V)SI/N - (\mu + \delta(V))I\,.\end{aligned} \tag{1.5}$$

Because of the separation of timescales, the model converges to a steady state before the appearance of a mutant. The number of susceptible individuals at this steady state equals $\bar{S}(V) = \frac{\delta(V)+\mu}{\beta(V)}$, which is identical to $1/\mathscr{R}_0(V)$. The question we now ask is whether a mutant with SPVL V' can invade a population infected by the resident strain with SPVL V. For this, we add another ODE, i.e., the number of patients with SPVL V' and linearize the equation at $I' = 0$. We find that the mutant can invade whenever

$$\frac{\beta(V')}{\delta(V')+\mu} > \frac{\beta(V)}{\delta(V)+\mu}$$

which we may write as $\mathscr{R}_0(V') > \mathscr{R}_0(V)$. Hence, the strain with the higher $\mathscr{R}_0$ "wins" or, equivalently, the strain with the lower $\bar{S}$ is able to invade. This equivalent description is known as the pessimization principle [162]. In Figure 1.2D, the pairs (V, V') for which invasion is possible are colored red. Each invasion results in a SPVL that is closer to the white dot in the middle, which is a global evolutionary stable strategy (ESS).

The major problem with the adaptive dynamics framework in the context of HIV-1 evolution is its main assumption, i.e. that mutations are rare. As the mutation rate of HIV is about 3.4×10^{-5} per nucleotide per replication cycle [141], this assumption may not be valid. Notice that a high mutation rate does not per se imply a high phenotypic mutation rate. However, as we will see below, the SPVL strongly depends on the genotype of HIV-1. That being said, adaptive dynamics can be used to create expectations, and has for instance been used to predict the evolution of HIV-1 virulence in situations where imperfect therapy and prophylaxis is applied [211].

Another framework that has been employed in order to understand SPVL evolution is the Price equation [PE; 177] framework. As a quick example, consider the following epidemiological model [cf. 42]. Let S denote the number of susceptible individuals, and let $I(v)$ be the density of individuals infected with a virus with $\log_{10}$ virus load v. The model is then given by the following ODEs:

$$\begin{aligned}\frac{\mathrm{d}S}{\mathrm{d}t} &= s - \mu S - S/N \int_v \beta(10^v)I(v)\,\mathrm{d}v \\ \frac{\mathrm{d}I(v)}{\mathrm{d}t} &= \beta(10^v)I(v)S/N - (\mu + \delta(10^v))I(v)\,.\end{aligned} \tag{1.6}$$

The PE describes the evolution of the average trait value $\bar{v}$ in terms of the covariance between the trait v and the Malthusian fitness $\beta(10^v)S/N - \delta(10^v)$:

$$\frac{\mathrm{d}\bar{v}}{\mathrm{d}t} = \mathrm{Cov}[v, \beta(10^v)S/N - \delta(10^v)]. \tag{1.7}$$

The PE thus predicts the direction and rate of the evolution of the mean trait $\bar{v}$. Notice, however, that the average and the covariance is taken with respect to the time-dependent distribution of v. In multiple populations it has been observed that the mean $\log_{10}$ SPVL is not constant over time. This has been interpreted as ongoing $\mathscr{R}_0$-optimization [90]. Recently, Blanquart *et al.* [19] applied the PE framework to SPVL evolution of HIV-1 in Uganda. Using the PE (together with deterministic and stochastic models), they predict stabilizing selection of the SPVL towards lower levels of virulence, which is in agreement with the observed attenuation of the virus.

Although it is possible to add effects of mutation to this model and use clever approximations to derive a more tractable version of the PE [19], the models that we will discuss in Chapters 4 and 5 are too complex to fit in the PE framework, as we incorporate multiple levels of selection. Instead, we use stochastic individual-based models (IBMs) to study the evolutionary dynamics.

1.3.2 Heritability of the SPVL

One necessary condition for adaptive SPVL evolution is that the SPVL is a heritable trait of the virus. A pathogen trait is heritable when the observable variation in the host population is at least partly due to genetic variation of the pathogen. The extent to which a quantitative trait x is heritable is measured by the (broad-sense) heritability statistic, denoted H^2, which is defined as

$$H^2 := \frac{\mathrm{Var}[g]}{\mathrm{Var}[x]}. \tag{1.8}$$

Here, we again write $x = g + e$, with g the genetic component of x and e the environmental component. We and many other authors often mistakenly write h^2 to denote the broad-sense heritability, and often omit the "broad-sense" qualifier. In fact, the symbol h^2 should be used for the "narrow-sense heritability", which is the largest fraction of the trait's variance that can be explained with linear regression on the alleles at the relevant quantitative trait loci. Hence, narrow-sense heritability does not measure epistatic interactions between loci.

As the quantity g is in general not observable directly, various statistical methods have been developed to measure H^2 by other means. Statistically, the simplest method involves pairs (or larger groups) of patients that are known to be infected with genet-

ically identical, or at least highly similar strains of the pathogen. For instance, one can consider the cherries of the pathogen's phylogeny [20, 205], or use transmission couples—pairs of patients of which one is known to have infected the other [98]. The heritability can be estimated from transmission couples with simple linear regression. Under the right assumptions, H^2 is equal to the slope of the regression line between the trait value of the transmitter and the receiver.

Another approach for estimating heritability uses (variations of) the so-called phylogenetic mixed model (PMM). In the PMM, the genetic component of a trait g evolves on the branches of a phylogenetic tree of the pathogen. The trait value at the tips of the tree is known, and in order to estimate H^2, one has to infer the rate of genetic drift. Evolution of g can be modeled as a Wiener process, but recent studies [18, 20, 152] have successfully used the Ornstein-Uhlenbeck process, which has the advantage over the Wiener process that it is stationary. In the case of HIV, a stationary process might be better suited, since the variance of the SPVL does not appear to increase with time, which would be a prediction of models that use a Wiener process.

When Fraser *et al.* [65] published their paper, there was only limited evidence of the heritability of the SPVL, and it was not very clear how large H^2 was, or how large it needed to be in order to explain the SPVL distribution. In the last decade, however, many studies have provided evidence for positive SPVL heritability [160], and analytical modeling shows that a high heritability provides the "most parsimonious" explanation for the large observed variance of the SPVL [22]. A recent reliable estimate is $H^2 \approx 30\%$, i.e. about one third of the variance of the SPVL is explained by viral genetic factors [20], although this figure does depend on HIV-1 clade and host population.

The presented estimators for H^2 are only reliable under the right conditions [152]. We stress this point in Chapter 4, where we use an IBM of the HIV-1 epidemic to show that HLA-polymorphism could result in an over-estimate of the heritability of the SPVL in the case of the transmission-couple method.

1.3.3 Modeling SPVL evolution

Several models have been developed for studying SPVL evolution. Shirreff *et al.* [206] construct a compartmentalized model with a discrete time step. The population variance σ_P^2 of the $\log_{10}$ SPVL is assumed to be the sum of 3 components,

$$\sigma_P^2 = \sigma_G^2 + \sigma_M^2 + \sigma_E^2, \tag{1.9}$$

where σ_G^2, σ_M^2, and σ_E^2 are respectively the genetic, mutational, and environmental components of the variance. In this model, the genetic component g of the $\log_{10}$ SPVL is allowed to mutate at the moment of transmission, according to a $\mathcal{N}(g, \sigma_M^2)$ distribution. The heritability is then given by $H^2 = \sigma_G^2/\sigma_P^2$. The next-generation matrix

(NGM) is employed to compute a mutation-selection balance, which can be used to estimate the model parameters σ_M and σ_E from transmission-couple data. Shirreff *et al.* [206] conclude that the evolutionary hypothesis [*sensu* 65] can be understood with a relatively simple model, and that SPVL evolution towards the $\mathscr{R}_0$-maximizing value is expected to happen within 1 or 2 centuries, which is reasonable considering the estimated origin of HIV-1 [239].

Similar conclusions can be reached with a stochastic IBM, which allows for more details such as a sexual-contact network [89]. The IBM of Herbeck *et al.* [89], can also be used to study the impact of widespread ART coverage on the evolution of SPVL. According to the model, although a high ART coverage decreases HIV-1 incidence, the mean $\log_{10}$ SPVL could increase from 4.7 to 5.1 $\log_{10}$ copies per ml within a couple of decades. This model assumes neutral mutation of the SPVL at the moment of transmission.

1.3.4 Short-sighted evolution

In the models cited above, the within-host evolution of the SPVL is relatively neutral. Both Bonhoeffer *et al.* [22] and Blanquart *et al.* [19] include a term that represents intra-host evolution of the virus-genetic component of the SPVL. When this term is positive, selection within the host favors strains that are more virulent, possibly due to a higher replicative capacity. When the intra-host evolution term is negative, within-host evolution results in an attenuated virus, possibly due to immune-escape mutations that induce a fitness cost.

The length of an untreated HIV-1 infection ranges from 2 to more than 10 years, a period in which the virus undergoes many cycles of replication. Although the transmission interval can be much shorter than the duration of the infection [225], HIV-1 is expected to experience much more selection for within-host fitness than between-host fitness (i.e. $\mathscr{R}_0$). Evolution that increases within-host fitness, but not between-host fitness is known as short-sighted evolution [126].

In order to address the influence of short-sighted evolution on HIV-1 SPVL evolution, Lythgoe *et al.* [139] developed a multi-level model that incorporates selection for intermediate SPVLs on the population level (using the infectiousness and virulence estimates of Fraser *et al.* [65]), while the individual-level model allows for competition between multiple strains using quasi-species dynamics, without explicitly modeling immune responses. The extent of within-host competition can be adjusted, and when this parameter is large, the mean $\log_{10}$ SPVL is easily skewed away from the $\mathscr{R}_0$-optimizing value from the neutral models. Notice that since $\mathscr{R}_0$ in such a model is defined for the entire quasi-species, optimization of the reproduction number is not as clearly defined as in the adaptive dynamics framework.

Host heterogeneity in the models discussed above is either absent, or merely comes in the form of additive noise (on the log scale). This means that within-host selection always acts in the same "direction", and can easily disrupt the effect of between-host selection that supposedly led to the observed SPVL distribution around $\mathscr{R}_0$-optimizing values. In Chapters 4 and 5 we incorporate a more complex type of host-heterogeneity in the form of HLA polymorphism. In our models, viruses adapt to their host by escaping from HLA-restricted immune responses, which means that the direction of within-host evolution is different for every host. We test the hypothesis [66] that this complex form of host heterogeneity allows us to reconcile within-host competition with population-level $\mathscr{R}_0$ optimization by means of SPVL evolution.

In Chapter 4, we construct a relatively simple IBM of HIV-1 immuno-epidemiology, that uses a phenomenological description of the within-host evolution of the virus in terms of immune escape and the reversion of deleterious mutations. The hosts are assumed to be maximally heterogeneous, which allows us to considerably simplify the representation of a virus, and make use of the NGM framework. Our model does not predict maximization of the population-level $\mathscr{R}_0$.

As the model developed in Chapter 4 is a bold simplification of reality, we use Chapter 5 to present a much more realistic HIV-1 model, both on the the individual and population level. The within-host model features coexisting HIV strains that compete for target cells, and can escape from multiple immune responses. We confirm some of the conclusions from Chapter 4: evolution remains short sighted.

1.3.5 The spread of CD8$^+$ T-cell escape mutations in populations

During an HIV-1 infection, the virus escapes from multiple CD8$^+$ T-cell responses. When such an escaped virus is transmitted, the receiver inherits the escape mutations that were acquired in the previous host environment [124]. This means that these escape mutations can potentially spread in the population and even go to fixation.

Since escape mutations are HLA-restricted, many strong associations between HLA type and polymorphisms in the genome of HIV-1 have been observed [154], which has been referred to as the "HLA footprint" on HIV-1 [147]. However, the spread of these escape mutations through the population can obfuscate these associations [122], and sophisticated statistical models are required to correct for the accumulation of these escape mutations [26].

The spread of escape mutations can have a real clinical impact. A transmitted virus that contains escape mutations restricted to one of the receivers HLA alleles is called pre-adapted [27] and pre-adaptation is associated with faster progression to AIDS. Conversely, many transmitted HLA-mismatched escape mutations are associated with slower disease progression [153]. This might be because escape mutations reduce the

viral replicative capacity in the average host, thereby reducing the SPVL and the rate of disease progression [173].

A small number of models for the spread of escape mutations of HIV-1 have been developed previously. For instance, Fryer *et al.* [69] developed a population-level ODE model with compartments for patients infected with HLA-matched and mismatched escape mutants, and used data from patients to infer the rates of escape and reversion. The same model can be used to explain biases in associations between HLA alleles and HIV-1 polymorphisms [68].

These models only consider a very small number of epitopes and HLA alleles. In Chapter 5 we use an IBM and massively increase the number of HLA alleles in the simulated population, as well as the number of possible epitopes in the virus. This allows us to simulate the HLA footprint on HIV-1, and the long-term effect of pre-adaptation on virulence.

Part I

An acute infection: influenza A virus

Chapter 2

Long-term adaptation of the influenza A virus by escaping cytotoxic T-cell recognition

Rutger G. Woolthuis[1,2], Christiaan H. van Dorp[1,2], Can Keşmir[1], Rob J. de Boer[1], and Michiel van Boven[2]

[1]Theoretical Biology, Utrecht University, Utrecht, The Netherlands.
[2]National Institute for Public Health and the Environment, Bilthoven, The Netherlands.

Nature Scientific Reports (2016)

Abstract. The evolutionary adaptation of the influenza A virus (IAV) to human antibodies is well characterized. Much less is known about the long-term evolution of cytotoxic T lymphocyte (CTL) epitopes, which are important antigens for clearance of infection. We construct an antigenic map of IAVs of all human subtypes using a compendium of 142 confirmed CTL epitopes, and show that IAV evolved gradually in the period 1932-2015, with infrequent antigenic jumps in the H3N2 subtype. Intriguingly, the number of CTL epitopes per virus decreases with more than one epitope per three years in the H3N2 subtype (from 84 epitopes per virus in 1968 to 64 in 2015), mostly attributed to the loss of HLA-B epitopes. We confirm these observations with epitope predictions. Our findings indicate that selection pressures imposed by CTL immunity shape the long-term evolution of IAV.

2.1 Introduction

IAV persists by continually escaping pre-existing immunity in the population. Most attention has been on the evolution of surface proteins hemagglutinin (HA) and neuraminidase (NA) that form the main targets of neutralizing antibodies [135, 196, 233]. Antibody mediated immunity is subtype specific and lasts for 2-7 years due to rapid evolution of the antigenic sites on the HA and NA proteins [12, 64, 210]. IAV also elicits CTL immune responses [116, 220], which reduce viral spread within the host by killing infected cells. As with memory B cells, memory CTLs mount a fast immune response upon recognition of epitopes years after the primary infection [117, 195], such that individuals with pre-existing CTLs develop less severe disease [148, 214]. CTLs also provide heterosubtypic immunity [179, 232], which could be an attractive feature for universal vaccines [115].

Viruses escape CTL recognition by mutating amino acid residues within CTL epitopes. Such immune escape mutations play an important role in the within-host dynamics of chronic pathogens (e.g. HIV-1) and are also observed during acute IAV infection [15, 176]. While immune escape mutations in IAV cripple the virus [16, 79], these mutations can persist in a prolonged infection [222], and at the population level despite the high polymorphism of the human leukocyte antigen (HLA) [77, 103]. Recently, positive selection in CTL epitopes has been shown in the nucleoprotein (NP) by comparing human and swine viruses in a phylogenetic analysis [140]. Many CTL epitopes have been identified in IAV [7, 25], but a framework capturing the dynamics of CTL epitopes in all proteins over long evolutionary time is lacking. Here we analyze historical and contemporary IAV sequence data spanning the period 1932-2015, using 142 empirically confirmed CTL epitopes known to date [21, 224] (Table S2.1, Table S2.2, and Table S2.3; Methods).

2.2 Results

2.2.1 Antigenic cartography based on CTL epitopes

We combine 295 representative human IAVs and the compendium of CTL epitopes into an antigenic map that tracks the long-term evolution of CTL epitopes in IAV across the H1N1, H2N2 and H3N2 subtypes (Figure 2.1). Each virus contains a subset of the CTL epitopes (Figure S2.1 and Figure S2.2), with on average 74 epitopes per virus (summed over all class I HLAs). In total, we find 134 out of the 142 epitopes in these viruses, of which 24 are conserved in the study period (marked in Table S2.2). At seven loci (positions in the proteome) we find more than one confirmed epitope, i.e. at these loci epitope variants have mutated at some point in time to another epitope variant.

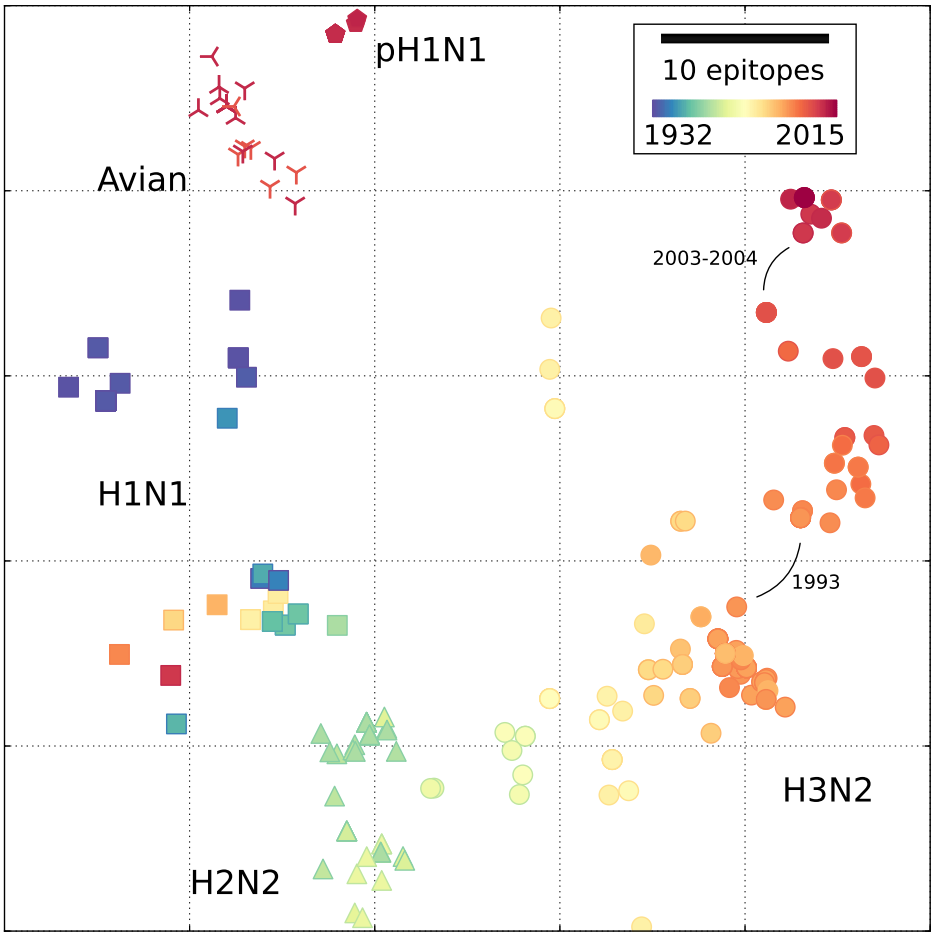


Figure 2.1: CTL epitope evolution in the influenza A virus. Antigenic map of 295 representative influenza A viruses spanning the period 1932-2015 (□ H1N1, △ H2N2, ◯ H3N2, ⬠ pH1N1) based on 134 CTL epitopes. The H3N2 subtype has evolved extensively over the period 1968-2015, while the H2N2 and H3N2 viruses circulating in the late 1960s are antigenically close. Recent avian viruses are superimposed independently onto the antigenic map (⅄ H5N1, ⅂ H7N9 and ≺ H9N2), using the 134 CTL epitopes of human IAV origin. The map is constructed using multi-dimensional scaling (MDS) based on Jaccard distances, explaining 93% of the antigenic distances (Methods). The scale bar denotes the expected difference in the number of epitopes; the colors indicate the collection year of the virus.

As a measure of immunological similarity we use the Jaccard index, defined as the number of epitopes shared by a pair of viruses divided by their number of unique epitopes (Methods). This measure is attractive biologically as it is based on overlaps of epitopes between viruses, and counts any mutation in an epitope as a CTL escape. Multidimensional scaling (MDS) based on Jaccard distances yields a map in which the distance between any pair of viruses represents the expected number of different epitopes (Figure 2.1). The map accurately visualizes the expected cross-immunity between viruses, even across subtypes ($R^2 = 0.93$, Figure S2.3). We find similar results using Manhattan distances (Figure S2.3). Similar maps based on antibody-mediated immunity can only be constructed for each subtype separately, because hemagglutination inhibition assays are subtype specific [12, 210].

In the CTL antigenic map, a time directional evolutionary path runs from the early A/1932 (H1N1) to the recent A/2015 (H3N2), with the H2N2 subtype located between the H1N1 and H3N2 subtypes (Figure 2.1). The highly reassorted 2009 pandemic H1N1 (pH1N1) virus is distinct from all other viruses, and is genetically (Figure S2.4) and antigenically (Figure 2.1) most closely related to the early twentieth century H1N1 viruses.

The early H3N2 viruses (1968-1970) are remarkably close to the previously circulating H2N2 viruses (1957-1967; Figure 2.1). Conversely, the H3N2 subtype has evolved by more than 35 CTL epitopes over the past 46 years (Figure 2.1), with noticeable antigenic jumps in 1993 and 2003-2004. The near-continuous H2N2-H3N2 transition results from the fact that only HA and polymerase basic 1 protein (PB1) were reassorted, while HA carries few known CTL epitopes and PB1 is a highly conserved protein. The antigenic jump in 1993 results from a R384G mutation in NP, affecting four epitopes that are restricted by the HLA alleles B*44, B*08:01, B*27:02, and B*27:05 [187]. In 2003-2004, five epitopes mutated in the NP, M1, and PA proteins. The CTL antigenic map differs completely from a map based on amino acid sequences of the viruses (Figure S2.4). Here subtypes are clearly separated, mostly due to the variable HA and NA proteins. Excluding HA and NA from the amino acid map decreases the difference with the epitope map, in particular with respect to the subtype clustering (Figure S2.4). Nevertheless, there are still important differences between the antigenic and genetic maps, like the absence of the antigenic jumps in the amino acid map. Overall, IAV gradually drifts away from ancestral viruses by escaping CTL epitopes.

Not all epitopes are expected to impose equal immune pressures on the IAV. First, epitopes vary in immunodominance due to heterogeneity in HLA-allele frequencies in the population. However, when weighting the epitopes with the corresponding HLA supertype frequencies of five main ethnicities, we find that the distances between viruses are hardly affected (Figure S2.5). Similarly, some epitopes are intrinsically immunodominant by evoking stronger immune responses than others, e.g., due to differences in precursor frequencies [112]. Since the relative immunodominance of

epitopes is not known, weighting for their exact contribution is impossible. However, when we assign random weights to the relative contribution of the different epitopes, we find that even with substantial immunodominance the results are unaffected (Figure S2.6). Thus, while individual immune-responses might be affected by dominant CTL epitope escapes, the analyses show that at the population level the pattern of viral CTL epitope evolution is robust.

The antigenic cartography allows us to investigate avian IAVs that cause human infections and may pose a pandemic threat [91, 179, 231]. To characterise the antigenic relatedness of avian viruses to human viruses in the context of human CTL immunity, we superimpose representative avian viruses of the subtypes H5N1, H7N9 and H9N2 onto the antigenic map (Figure 2.1; Methods; Table S2.4). The avian viruses are more similar to pH1N1 viruses than to recent H3N2 viruses, suggesting that prior infection with a pH1N1-like virus is likely to yield more CTL immunity to the avian viruses than infection with recent H3N2 viruses. This illustrates that positioning of a virus in the antigenic map is possible on the basis of its sequence, enabling assessment of the extent of pre-existing CTL immunity in the human population.

2.2.2 Escape from CTL mediated immunity

We hypothesise that IAV evolution is in part driven by adaptation to pre-existing CTL immunity. To test this hypothesis, we compile a data set of 62 HLA binding peptides of IAV origin that do not elicit a CTL response (Methods; Table S2.3). With the exception of a small number of early H1N1 viruses in the period 1932-1935, the confirmed epitopes are significantly more variable than the non-immunogenic HLA binding peptides (Figure S2.7 and Figure S2.8). Additionally, the viruses are located considerably closer together in a map obtained using distances based on the non-immunogenic peptides than in the antigenic map (Figure S2.9). The difference in the evolutionary rates between CTL epitopes and non-immunogenic peptides suggests that there is selection for escape of CTL epitopes.

To substantiate this further, we use 7 347 whole-proteome sequenced viruses (Methods), and find that the total number of epitopes per virus decreases over time (Figure 2.2A). In H3N2 approximately one epitope is lost every two-and-a-half years, and in H1N1 approximately one epitope is lost every eight years. Since these time series are ancestrally dependent, we also consider the annual (relative) change in the number of epitopes (Methods), and reject neutral evolution for H3N2 ($p = 0.043$, Wilcoxon signed-rank test). We cannot distinguish neutral from directed evolution for H1N1 and H2N2, perhaps because the decrease is too small for H1N1 and the time series is too short for H2N2. The number of epitopes decreases most markedly in NP, while remaining almost constant in PB1 (Figure 2.2B and Figure S2.10).

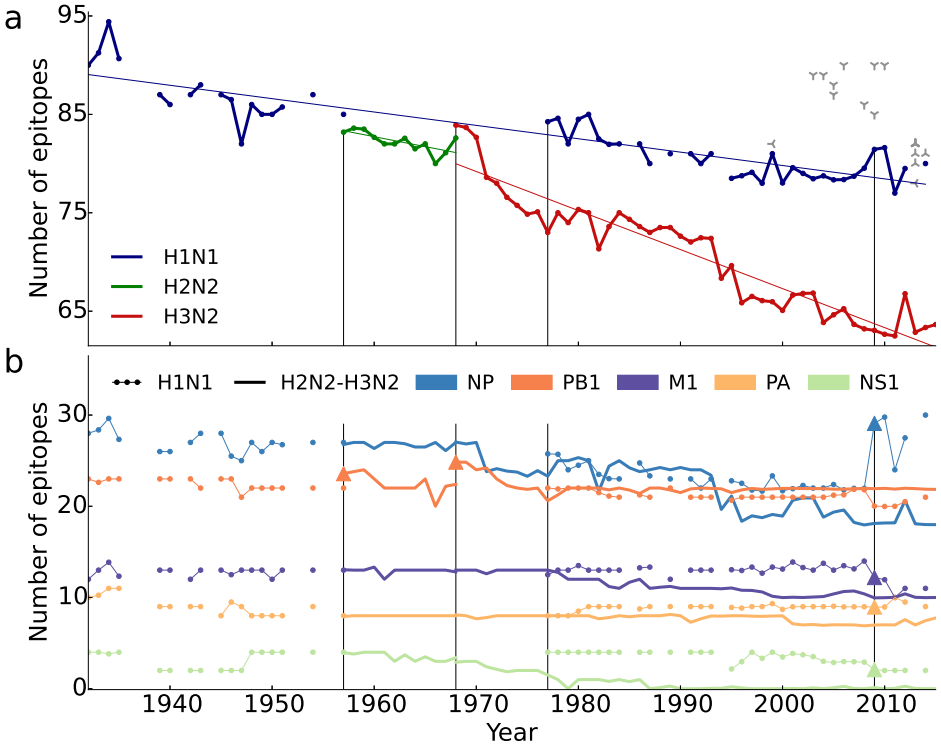


Figure 2.2: Decrease in the number of CTL epitopes over time. A. The number of CTL epitopes per virus decreases with 0.14 ± 0.01, 0.20 ± 0.06, and 0.40 ± 0.02, epitopes per year ($\pm$ SEM) in H1N1, H2N2, and H3N2, respectively (Methods). The number of CTL epitopes in avian viruses is larger than in recent human viruses (⅄ H5N1, ⅂ H7N9 and ≺ H9N2). **B.** Protein specific CTL epitope dynamics showing a decrease in the number of epitopes in NP, matrix 1 protein (M1) and non-structural protein (NS1). The other IAV proteins are laid down in Figure S2.10. Reassortment events resulting in the appearance of novel proteins are marked by triangles. Subtype replacements are marked by vertical lines. The analysis is based on 7 347 whole-proteome sequenced viruses (Methods).

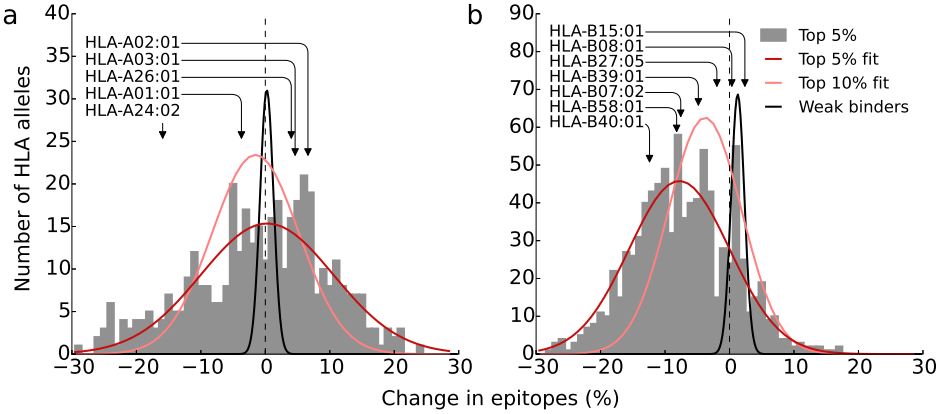


Figure 2.3: Decrease in the number of predicted epitopes. The change in number of top 5% strongest binding peptides in NP of H3N2 viruses for HLA-A (**A**) and HLA-B (**B**) alleles is shown over the period 1968-2015 (histograms; Methods). These peptides decrease in number for the large majority of HLA-B alleles ($p = 0.006$; mean change $\hat{\mu} = -9.1\%$ (95% CI: $[-15.6, -2.6]\%$) over the study period, Methods), but not for HLA-B alleles ($p = 0.46$). Despite the high level of heterogeneity in binding motifs, we find a decrease in the number of top-binders for a broad range of HLA-B alleles (Figure S2.12). The position of the 12 HLA supertype representatives are indicated with arrows. Fitted Gaussian curves indicate the decrease calculated for the thresholds $\leq 5\%$, $\leq 10\%$ (top binders), and $> 50\%$ (non-binders).

The decrease in the number of CTL epitopes could possibly be due to an underrepresentation of recently emerged epitopes, for which the identification could be less likely than for CTL epitopes in old viruses. However, the number of non-immunogenic peptides is constant over time (Figure S2.11), suggesting that the decrease is unlikely the result of observational bias.

In addition, we obtain an independent test by computationally predicting potential CTL epitopes. These predictions are free from any observational bias that may be present in the set of empirical epitopes. In the analysis we use prediction tools based on HLA-peptide binding affinities [99], using a suite of binding thresholds and focusing on the top 5% binders. With this threshold 88% of the empirical epitopes is present in the predicted set (Figure S2.12 and Figure S2.14). For NP in H3N2, which harbors most of the empirical epitopes (Figure 2.2B), we find a decrease in the number of strongly binding peptides for the large majority of HLA-B alleles (Figure 2.3). We find a similar though non-significant trend when combining the NP, M1, and PB1 proteins in H3N2 (Figure S2.14). These observations are in agreement with a significant decrease in the number of empirical epitopes restricted to HLA-B alleles (Figure S2.15). The decrease is two-fold larger in HLA-B restricted epitopes compared to HLA-A restricted

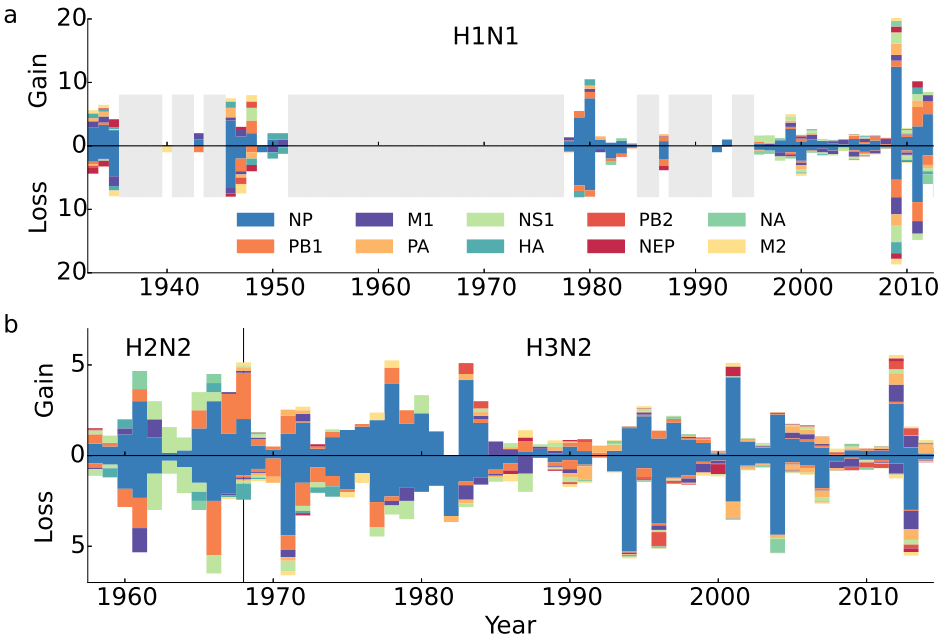


Figure 2.4: CTL epitope dynamics. Epitopes appear and disappear continuously in H1N1 (**A**) and H2N2-H3N2 (**B**), with the highest turnover in NP. Within years epitope gains and losses are correlated, due to annual variation in genetic drift. We calculate for each epitope the difference in the fraction of viruses harboring the epitope in a given year and the year before. Summing over all positive (negative) differences of the fractions gives the total gain (loss).

epitopes. The predictions show that the decrease in the number of empirical epitopes per virus (Figure 2.2A) is not restricted to the empirical set of epitopes.

The loss of CTL epitopes is the long-term net result of a continuous and rapid turnover of epitopes (Figure 2.4). In the short-term, genetic drift of the virus results in similar (and correlated) losses and gains of epitopes (Figure 2.4). Here we calculated the losses and gains of the epitopes weighted by their frequency in the 7 347 viruses. Of the two proteins containing most epitopes, NP has a higher turnover of epitopes than the more conserved PB1, in H1N1 as well as in H3N2. As in the antigenic map we find no unusual losses or gains during the H2N2-H3N2 transition. The turnover of epitopes shows that only a fraction of the escapes results in a lower number of epitopes per virus.

2.3 Discussion

By combining CTL epitope data with IAV sequence data, we provide evidence that IAV evolves to escape preexisting cellular immunity in the population. Namely, mutations accumulate more rapidly in epitopes than in non-immunogenic peptides, and the total number of epitopes decreases over time (Figure S2.7 and Figure 2.2).

Mutations in epitope sites can result in two different types of escape. First, T-cell receptor (TCR) escapes prevent memory T cells from recognizing the mutated epitope. In this case the virus circumvents preexisting memory responses against the epitope in the population, but can trigger the expansion of other immune responses upon infection. The second type are escapes that prevent loading of the epitope. This happens when the mutation either prevents HLA-binding or the processing of the epitope. In this case the host cannot elicit any T-cell response against the epitope. Combinations of these escapes are also possible when the mutation lowers the TCR as well as the HLA binding affinity.

In our analysis we treat both escapes equally, since the exact impact of the mutations is unknown. This might be an oversimplification, because some of the mutations may hardly affect the binding with either TCR or HLA. However, experiments have shown that CTL escape mutations usually affect the immune response [15, 16, 176, 187]. In addition, the decrease over time in the number of strongly binding predicted peptides indicates that the overall HLA binding affinity decreases as well, confirming that mutations in epitopes lower the HLA binding. Therefore, we believe that most mutations do affect the CTL response and argue that the simplification of treating any mutation in epitopes as escape is justified.

The adaptation to preexisting CTL immunity that we observe is in agreement with a recent study showing positive selection at CTL epitope sites in NP by a comparison of evolutionary rates in swine and human IAVs [140]. Using a phylogenetic analysis the authors find an increased rate of epitope mutations on the trunk of the tree in human compared to swine IAVs. In line with this, we find that the change in epitopes is larger than the change in non-immunogenic peptides (Figure S2.7). In addition, we find that CTL epitope loci disappear over time, which is a direct indication of adaptive evolution (Figure 2.2 and Figure 2.3). In agreement with Machkovech *et al.* [140], the decrease in the number of CTL epitopes is most pronounced in NP, although we see similar trends in other proteins.

At this stage the impact of the lower number of CTL epitopes in IAV is unknown, as there are no studies done showing changes in the virulence of influenza. Correlating incidence data with epitope escapes is complex, since other factors, like vaccination and the rapid antigenic drift due to neutralizing antibodies [12], interfere with the CTL mediated drift. While additional experiments have to unravel the phenotypic effect

of the adaptation, circumventing preexisting CTL immunity could in principle affect the severity of disease, since the lack of preexisting cellular immunity results in more severe infection [148, 214, 236].

We speculate that the decrease in the number of CTL epitopes has been ongoing since the global selective sweep of the internal proteins of IAV in the late 1800s [240]. Possibly, the number of CTL epitopes could return to a high value when zoonotic rearrangements introduce proteins that have not undergone evolutionary adaptation to human cellular immunity [140]. Examples are the increase of the number of epitopes in PB1 of H3N2 viruses in 1968, and the increase of the number of epitopes found in NP in the pandemic H1N1 viruses in 2009 (Figure 2.2B). Underpinning this further, we find a high number of epitopes in the avian viruses, especially in H5N1 (Figure 2.2A).

Escaping humoral immunity releases IAV from antibody-mediated selection pressures generated by viruses circulating in the recent past [12, 210]. This is not the case for cellular immune responses that have the potential to impact the long-term evolution of the virus. Since human IAVs seem to be adapted to CTL immune responses, novel viruses with unadapted reassorted internal genes would be at a selective disadvantage in the human population. This may explain why emerging IAVs with reassorted internal genes historically have only been successful when introduced in viruses that have also reassorted their HA or NA genes.

2.4 Methods

Complete virus set. We collected human IAV sequences (H1N1, H2N2 and H3N2) from the GISAID EpiFlu database from the period 1932-2015 [`www.gisaid.org`; July 2015; 21]. We selected viruses for which all protein sequences (except PB1-F2) are available and of which the length of each protein was not shorter than 20 amino acids of the consensus protein length. Moreover, we excluded any virus with sequences containing ambiguous amino acids (X, J, B). We excluded all H1N1 viruses collected in the year 1976, because these viruses originated from an exceptional swine flu outbreak in a military basis in New Jersey [78]. Moreover, A/Victoria/36/1988 was excluded, because this virus is a rare reassorted virus, and A/Canada/720/2005 (H2N2), because the subtype H2N2 did not circulate in 2005. This procedure resulted in 3 050 H1N1 (including pH1N1), 65 H2N2 and 4 213 H3N2 viruses.

Characteristic virus set. For computational reasons we compiled a subset with characteristic viruses spanning the period 1932-2015, and used this set for constructing the antigenic map. For the H1N1 and H3N2 subtypes, we included all viruses used by Westgeest *et al.* [233] and Bedford *et al.* [12], together with

A/Michigan/14/2014 (H3N2), A/Pennsylvania/44/2014 (H3N2), A/Sweden/1/2015 (H3N2) and A/Michigan/02/2015 (H3N2) to cover recent years, and included H2N2 viruses. In total, the characteristic set contains 295 viruses (Table S2.1), of which 48 are H1N1, 52 are H2N2, 185 are H3N2 and 10 are pH1N1.

Epitope compendium. We downloaded the epitope data from IEDB by using the filters "human" as host organism and "Influenza A virus" as source organism and selecting "Linear epitopes", "MHC Class I" and "T Cell Assays" [`www.iedb.org`; September 2015; 224]. Next, we gather unique epitopes with all corresponding immune assays. Some epitopes are embedded in longer epitopes belonging to the same HLA supertype [208]. In these cases, we kept only the embedded (shorter) epitopes. Thereafter, we removed epitopes with length larger than 12 amino acids, because the exact (shorter) epitope within these longer peptides are not reported. We also excluded epitopes that were not present in any of the IAVs in the complete set described above. Finally, we characterized the candidate epitopes either as immunological epitopes or as non-immunogenic peptides. The candidate epitopes with at least one reported non-negative T-cell assay are added to our epitope compendium, while candidates having negative T-cell responses in all assays are classified as non-immunogenic peptides. The list of 142 epitopes and 62 non-immunogenic peptides is provided in Table S2.2 and Table S2.3, respectively.

Jaccard distance. As a measure for the antigenic distance between two viruses we use a scaled adjusted Jaccard distance, reflecting the expected number of different epitopes between viruses i and j:

$$d_{ij} = \left(1 - \frac{|S_i \cap S_j|}{|T_i \cup T_j|}\right) \mathbb{E}[|T_k \cup T_l|]\,. \tag{2.1}$$

Here S_i and S_j represent the epitopes in viruses i and j, with $|S_i \cap S_j|$ the number of epitopes that two viruses i and j have in common. T_i and T_j represent the loci of viruses i and j carrying a functional epitope, with $|T_i \cup T_j|$ the number of unique loci in viruses i and j. Furthermore, $\mathbb{E}[|T_k \cup T_l|] = (1/N^2)\sum_{k,l} |T_k \cup T_l|$ is the expected number of unique epitope loci, with N the total number of viruses. To account for the fact that two different alleles at the same locus is a unit difference in epitopes, we deviate from the classical Jaccard distance by dividing by $|T_i \cup T_j|$ instead of $|S_i \cup S_j|$. Notice that $|S_i| = |T_i|$, because a virus can at most have one epitope at a given locus, and that the difference between the classical Jaccard distance, e.g., as used elsewhere [159], and our distance is small, since only a few epitopes in our compendium contain multiple alleles.

Manhattan distance. As an alternative to the Jaccard distance, we computed the Manhattan distance by counting the fraction of all possible epitope loci differing between pairs of viruses. In contrast to the Jaccard distance, this measure also counts epitope loci that are absent in both viruses, and therefore depends on the size of the epitope compendium. Additionally, the Jaccard index better counts the number of epitopes that would be recognized when encountering a second virus after infection with the first virus of a pair. Therefore, we believe that the Manhattan distance is less appropriate immunologically. Nevertheless, we found that there is an almost one to one relation between the two distance measures (Figure S2.3), such that the results do not change qualitatively, and to a large extent quantitatively, when using the Manhattan distance.

Pairwise proteome distance. We concatenated the protein sequences for each virus and perform a global alignment using the Clustal Omega package [209]. The whole-proteome pair-wise distances of the aligned viruses are calculated with Clustal Omega, which uses the k-tuple distance metric (with values between 0 and 1).

Constructing the antigenic map with multidimensional scaling. We follow Smith *et al.* [210] by using metric multidimensional scaling (MDS) to represent N viruses as elements of a low dimensional Euclidean space ($\mathbb{R}^n$), while preserving the distances between viruses as much as possible. Positions $x_1, \ldots, x_N \in \mathbb{R}^n$ for the N viruses are calculated by minimising the objective function E:

$$E(x_1, \ldots, x_N) := \sum_{1 \le i < j \le N} \left(d_{ij} - \|x_i - x_j\|_2 \right)^2 . \tag{2.2}$$

Here $\|\cdot\|_2$ is the Euclidean norm on the low dimensional space $\mathbb{R}^n$. We minimized the error for different initial configurations to ensure that the global minimum is found. With increasing target dimension n, the error decreases strongly and the variance in the distances captured (R^2) increases (Figure S2.3).

Superimposing avian viruses. From GISAID, we collected 21 avian viruses of the H5N1, H7N9, and H9N2 subtypes. To assess the extent of human cellular immunity against avian viruses, we determine their position in the antigenic map given the positions of the human viruses, i.e., we minimize E in Equation (2.2) for each of the 21 avian viruses independently. See Table S2.4 for an overview of CTL epitopes in the three avian subtypes causing human infections.

Predictions. In the HLA-peptide binding prediction analyses, we use all unique HLA alleles available in the NetMHCpan 2.8 prediction tool [99], and measure the relative

change in the number of top-binders over a period of time. We focus on the top 5% and the top 10% strongest binders, and use the $> 50\%$ (non-binders) as comparative control (the percentage is relative to a large set of peptides used in NetMHCpan). Most empirical epitopes belong to the 5% top binders (Figure S2.12).

Statistics. The slopes of the decrease in number of epitopes are calculated by linear regression (Figure 2.2A). Since the number of epitopes depends on the number of epitopes in previous years, we also calculated the difference in the number of epitopes from year-to-year (excluding missing years) and used the conservative Wilcoxon signed-rank test (one tailed) to estimate the significance of the decrease. The analysis with non-immunogenic peptides was performed similarly (Figure S2.11).

Since we expected the change in the number of binders ($\vec{x}$) to be alike for HLA alleles with similar binding motifs, we used a statistical model that corrects for these dependencies when determining the significance of the mean decrease per HLA allele. We assume that $\vec{x} \sim \mathcal{N}(\vec{\mu}, \Sigma)$, where $\vec{\mu} = (\mu, \ldots, \mu)^T$, and Σ is the covariance matrix describing the above-mentioned dependencies. We calculate Σ using the similarities between the binding motifs as follows: a HLA allele i can be represented as a binary string b^i indexed by all 9-mers k so that $b^i_k = 1$ if k is predicted to bind i, and 0 otherwise. We then choose $\Sigma_{ij} = s \cdot \mathrm{Cov}(b^i, b^j)$ for some constant $s > 0$, and obtain maximum-likelihood estimates for the mean decrease μ and scaling parameter s. This model is selected in favor of a model that assumes independence of the HLA alleles ($p < 10^{-12}$, likelihood-ratio test). The maximum likelihood estimates of the mean decrease and corresponding 95% confidence intervals are given in Figure 2.3.

2.5 Supplementary material

The supplementary tables are available at `www.nature.com/articles/srep33334` and contain:

Table S2.1: Characteristic influenza A virus set.
Table S2.2: Cytotoxic T lymphocyte epitopes in human influenza A viruses.
Table S2.3: Non-immunogenic peptides in human influenza A viruses.
Table S2.4: Human CTL epitopes in avian IAVs.

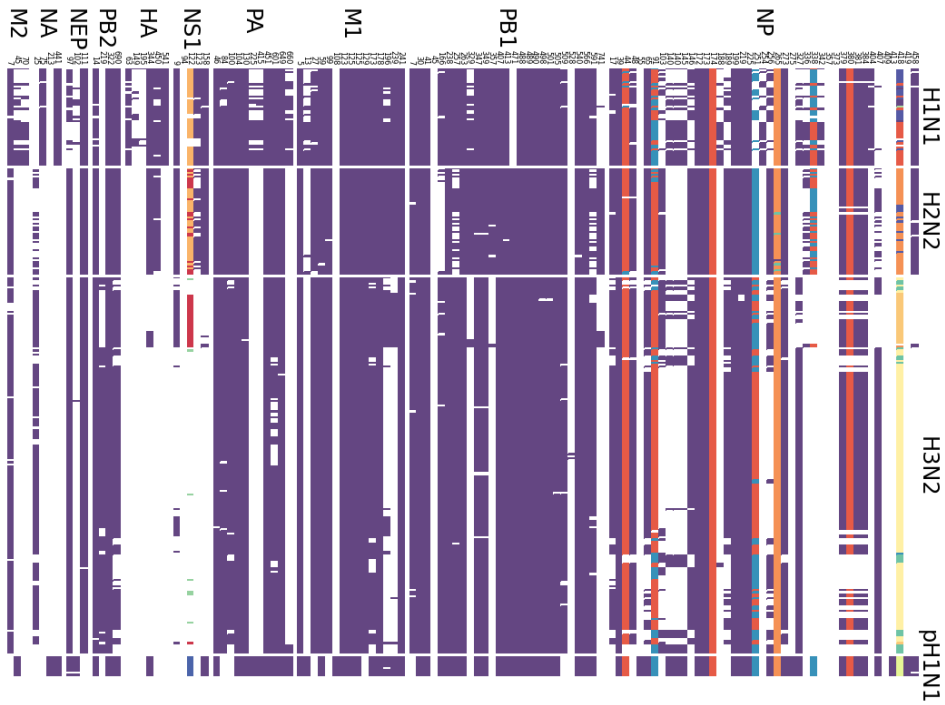


Figure S2.1: Overview of CTL epitopes by subtype. Colored rows indicate the presence of epitopes (horizontally ordered by protein and protein position) in the 295 viruses of the characteristic set (vertically ordered by subtype and collection year), using different (arbitrarily chosen) colors if multiple epitope variants are present at a locus (white: no epitope present). Single mutations can result in multiple epitope escapes, as observed in the NP_{379}, NP_{380}, NP_{381} and NP_{384} epitopes (subscript indicates the starting position of the epitope). Some epitopes re-emerge after a period of absence, e.g. the NP_{91} and NP_{251} epitopes. NP_{418} is the most variable epitope.

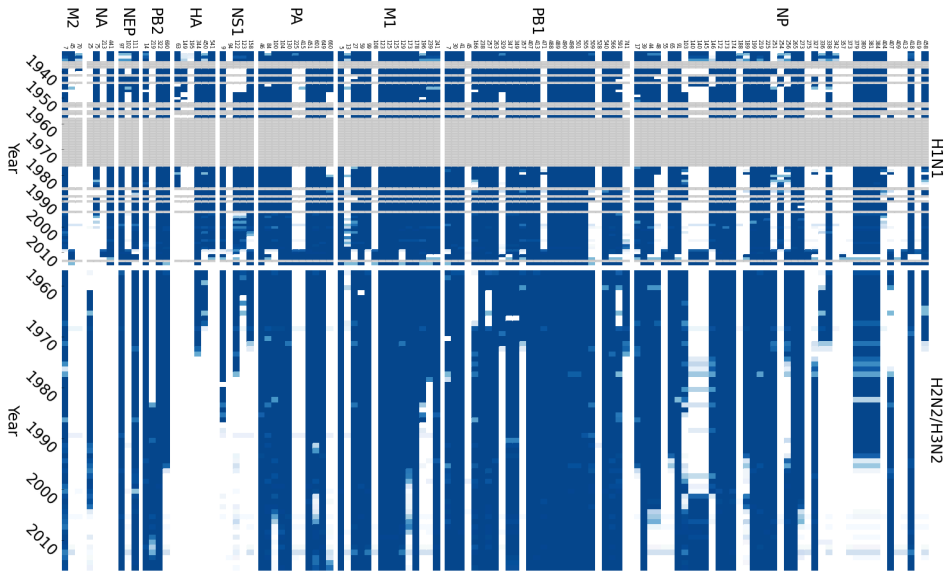


Figure S2.2: Overview of CTL epitopes by year. The fraction of viruses harboring the CTL epitope (ordered horizontally) in given year (ordered vertically; white: epitope not present; gray: no data available). H1N1 and H3N2 are plotted separately, because these subtypes have co-circulated since 1977. We use all whole-proteome sequenced viruses available in the GISAID EpiFlu database (3 050, 65, and 4 213 viruses for H1N1, H2N2, and H3N2 respectively).

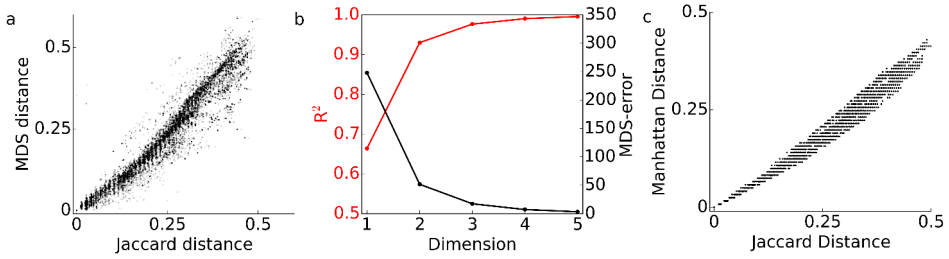


Figure S2.3: Variance explained by the antigenic map and comparison with Manhattan distance. A. The antigenic map of Figure 2.1 accurately visualizes antigenic distances, since the two-dimensional projected MDS pairwise distances capture most of the variance in the multidimensional Jaccard distances ($R^2 > 0.93$). **B.** R^2 saturates quickly with increasing the dimension of the target distance space. In addition, the minimized MDS error (Methods) is small already at low target dimension. **C.** Manhattan distance is highly correlated with Jaccard distance, yielding similar antigenic maps (not shown).

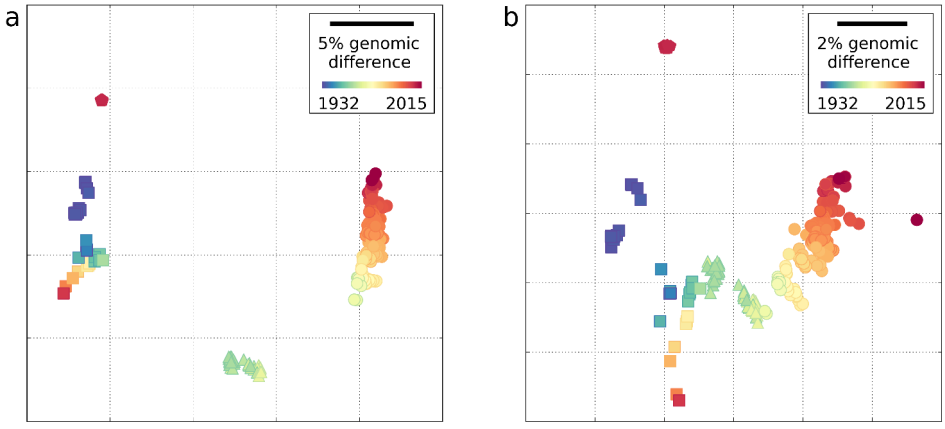


Figure S2.4: MDS maps of IAV proteomes. A. The geometry of proteome map differs considerably from the CTL antigenic map, with clear separation of subtypes (Figure 2.1 and Methods). $R^2 = 0.99$. **B.** Excluding the variable HA and NA proteins from the analysis brings the subtypes closer together, but the map remains different from the map based on epitopes. $R^2 = 0.97$.

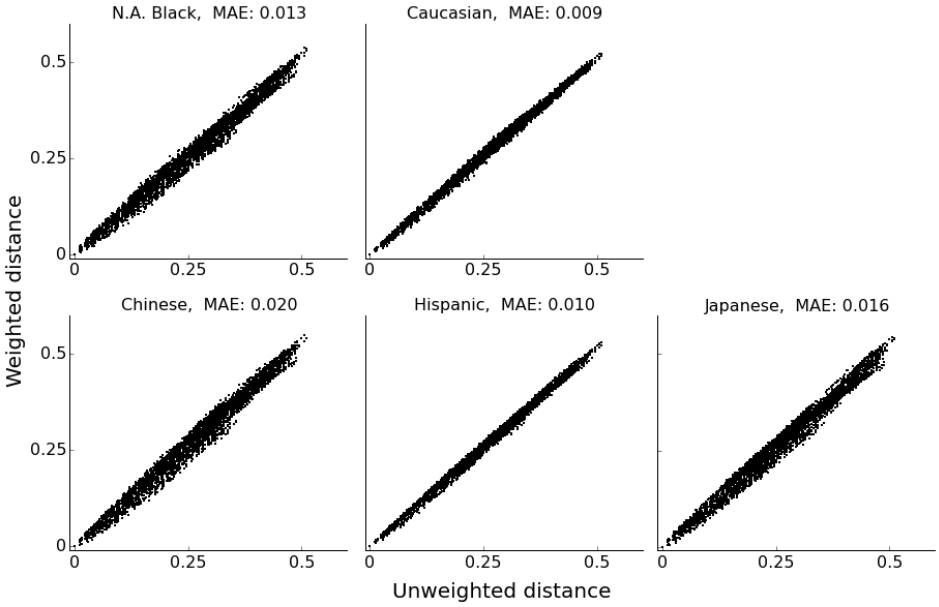


Figure S2.5: Antigenic distances weighted with HLA frequencies of main ethnicities. To test the robustness of the antigenic distance analysis to population immunodominance, we weighted the contribution of each epitope by the corresponding HLA supertype frequency in five main ethnicities (Black, Caucasian, Chinese, Hispanic, and Japanese) [203]. HLA supertype weighting has a minor effect on the pairwise virus distances for all ethnicities. This indicates that the epitope data set covers worldwide frequently occurring HLAs. MAE: mean absolute error compared to the diagonal.

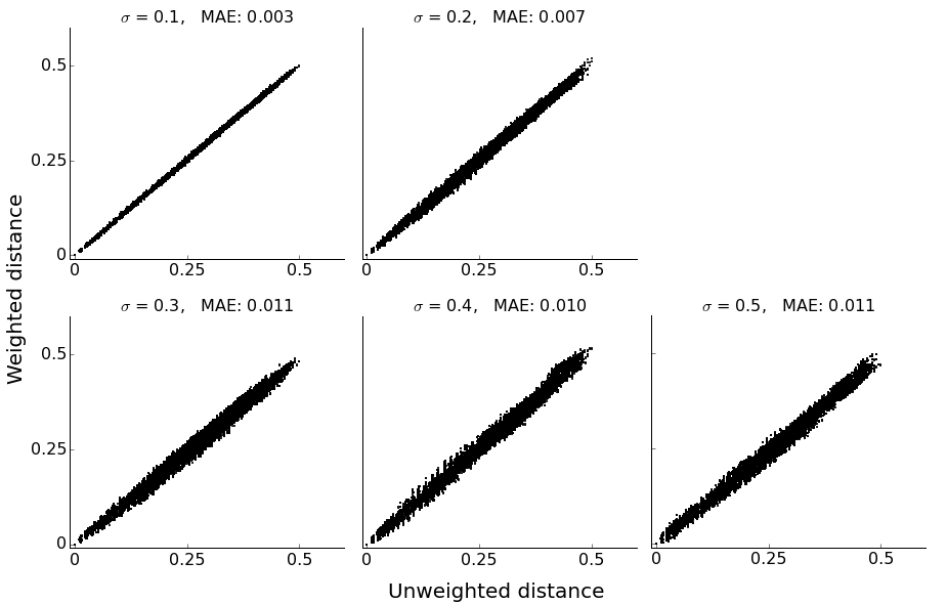


Figure S2.6: Robustness of viral distances to immunodominance. The relative immunodominance of individual epitopes is unknown. To test how variability in the relative immunodominance influences the antigenic map, we used a random weighting for each epitope by sampling from a (truncated) normal distribution with unit mean and five different standard deviations (σ), and compared weighted and unweighted distances. Even for large variability ($\sigma = 0.5$) we obtain only a small difference between weighted and unweighted distances (mean absolute error MAE $= 0.011$). This indicates that without knowing the exact relative immunodominance of the epitopes, the overall evolutionary pattern is accurately obtained using unweighted distances.

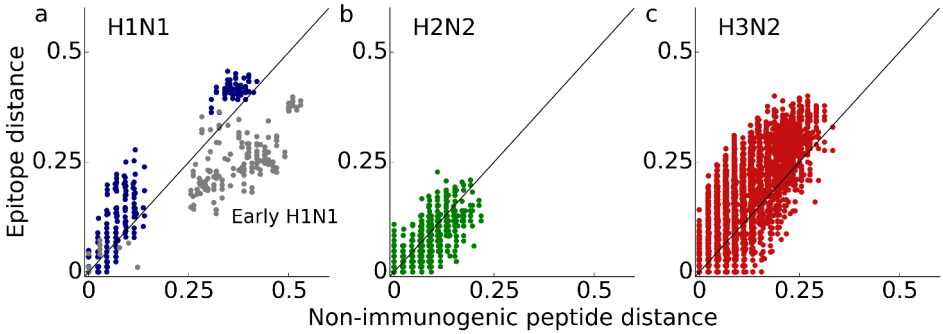


Figure S2.7: Evolution of CTL epitopes versus non-immunogenic peptides. The distances between all viruses based on CTL epitopes are larger than the distances based upon non-immunogenic HLA-binding peptides. We obtained the (normalized) pairwise virus distances of H1N1 (**A**), H2N2 (**B**), and H3N2 (**C**) viruses using 134 CTL epitopes and 62 nonimmunogenic peptides (Methods). A small number of H1N1 viruses from the years 1932-1935 differ more in their non-immunogenic peptides than in their CTL epitopes (grey dots). The H2N2 subtype circulated too shortly (1957-1968) to find a significant difference between epitope and non-immunogenic peptide distances. See Figure S2.8 for a statistical analysis of the comparison in H3N2.

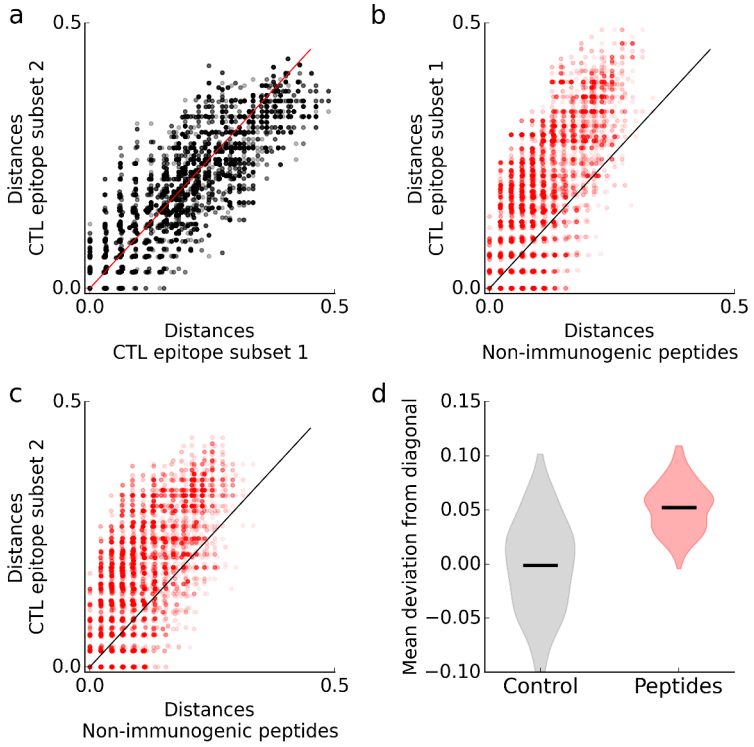


Figure S2.8: Statistical analysis of the evolution of CTL epitopes versus nonimmunogenic peptides in H3N2. To show that CTL epitopes evolve faster than non-immunogenic peptides, i.e. distances between viruses based on CTL epitopes are on average larger than those based on non-immunogenic peptides, we randomly divide the set of CTL epitopes into two equally sized subsets with as many epitopes as the number of non-immunogenic peptides (i.e. 62). We compare the CTL epitope subsets with each other (**A**), and with the non-immunogenic peptides (**B**, **C**), and find that distances based on the CTL epitope subsets are larger than for the non-immunogenic peptides. **D.** Repeating this analysis 100 times, we find a significant difference between comparisons among the subsets (control) and the comparisons with the subsets and the set of non-immunogenic peptides (peptides), measured as the average difference between each data point and the diagonal ($p < 10^{-5}$, Mann-Whitney U-test). Notice that not every epitope is used in each sample, as the number of CTL epitopes (134) is more than twice the size of non-immunogenic peptide set (62).

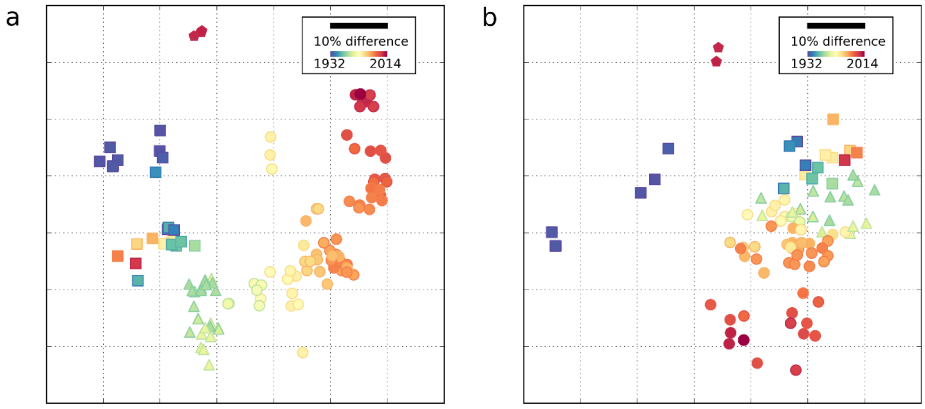


Figure S2.9: MDS map of non-immunogenic peptides. A. Antigenic map as in Figure 2.1 constructed using the confirmed CTL epitope compendium using unscaled distances (Methods). **B.** The map constructed using pairwise virus distances obtained using 62 non-immunogenic peptides contrasts with the antigenic map in panel **A**. In the map of panel **B**, most viruses are close, with the exception of early H1N1 and recent pH1N1 viruses. Only the early H1N1 viruses (isolated before 1935) and recent pH1N1 differ substantially due to reassortments. The non-immunogenic peptides do not elicit an immune response in any assay performed (Methods). The two-dimensional antigenic map in panel **B** explains most of the variance ($R^2 = 0.96$).

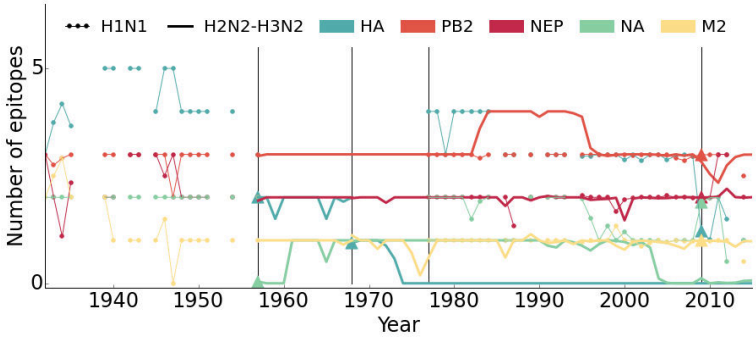


Figure S2.10: CTL epitopes in IAV proteins containing small number of epitopes. Time series of the number of CTL epitopes in proteins that harbor at most five epitopes. The number of epitopes in HA and NA decreases over time. Triangles mark reassortment events. See Figure 2.2B for the number of CTL epitopes over time in the other proteins.

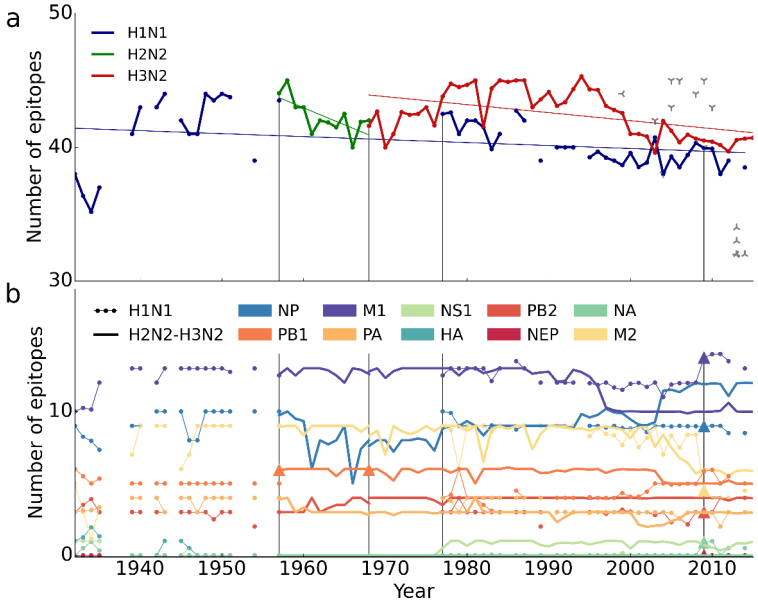


Figure S2.11: Dynamics of non-immunogenic peptides. A. No significant adaptive decrease is observed in the number of peptides per virus that are not immunogenic in H1N1, H2N2, and H3N2 ($p = 0.45$, 0.32, and 0.49, respectively; Methods). The decrease is 0.022 ± 0.011, 0.25 ± 0.08, and 0.06 ± 0.02 peptides per year ($\pm$SEM) for H1N1, H2N2, and H3N2, respectively. **B.** The number of non-immunogenic peptides per virus for each protein is constant over time, with only small changes in NP, M1, and M2 in the past two decades. We use peptides of IAV origin that are empirically tested and have a negative CTL response in all assays performed, and compute the occurrence in the complete virus set (Methods).

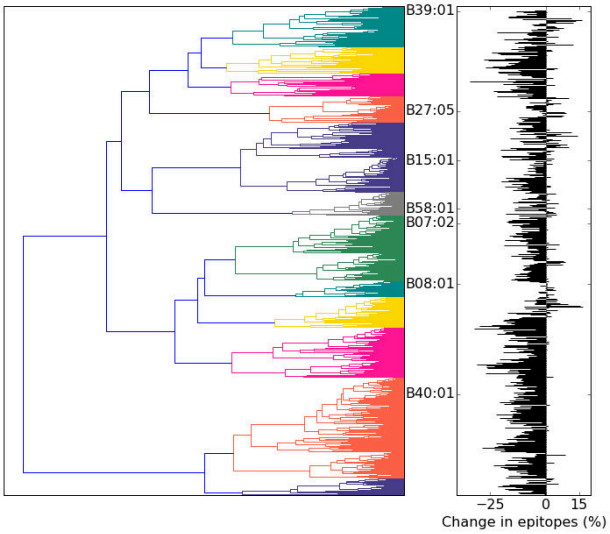


Figure S2.12: Prediction-based cluster analysis of HLA-B alleles. We find a decrease in the number of predicted epitopes [using NetMHCpan 2.8; 99] for the full HLA spectrum, indicating that a decreasing trend exists for HLA alleles with distinct binding motifs. The dendrogram (left) is based on Jaccard distances between HLA-B alleles using the 5% top binders in NP of H3N2 viruses. For each HLA allele the change in the number of predicted epitopes (top 5%) in H3N2 in the period 1968-2015 is indicated on the right. We labeled the 7 HLA-B supertype representatives, and set the cluster threshold such that no two supertype representative are in the same cluster. The dendrogram is constructed using the UPGMA algorithm.

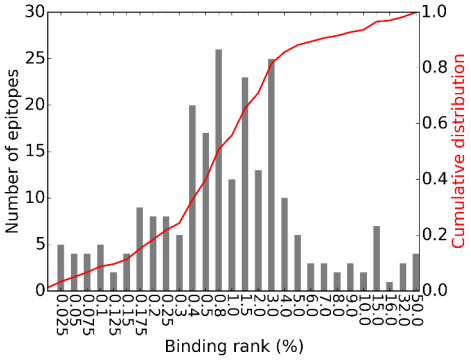


Figure S2.13: Predicted HLA binding affinity of the empirical epitopes. We use the prediction tool NetMHCpan 2.8 [99] to predict the binding affinities of the epitopes obtained from IEDB, and plot the number of epitopes in various binding rank bins. The binding of all HLA-epitope combinations are considered, when multiple HLA alleles per epitope are reported in IEDB.

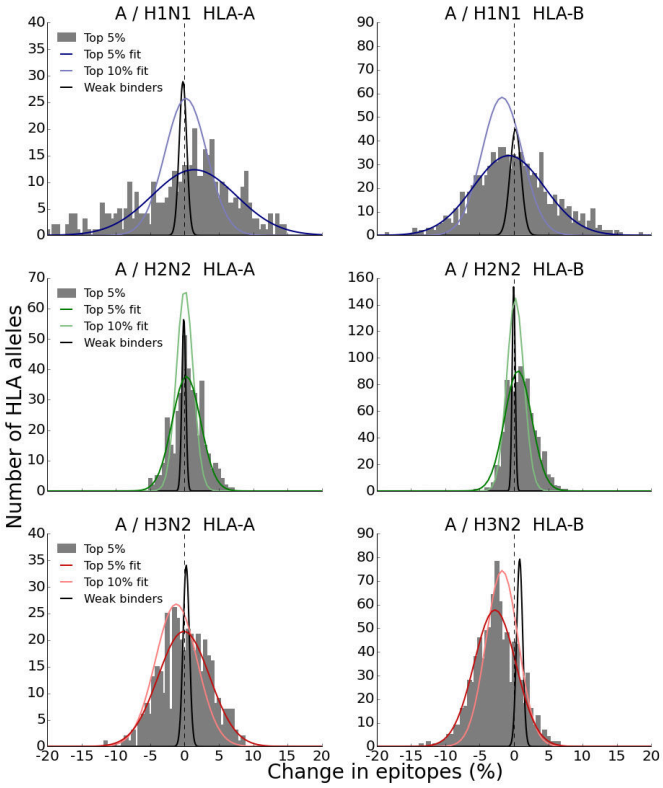


Figure S2.14: Change in predicted top binders in H1N1, H2N2, and H3N2. Aggregating the proteins NP, M1, and PB1 that cover most empirical epitopes (Figure 2.2B), we find a small decrease in the number of strongly binding peptides for most HLA-B alleles in the subtypes H3N2 ($p = 0.069$; mean change $\hat{\mu} = -3.6\%$ (95% CI: $[-7.5, 0.3]$%) over the study period). We find no decreasing trend in H1N1 and H2N2, probably because the latter subtype only circulated for a decade and because H1N1 evolves slower (Figure 2.2). The top 5% best binding peptides (relative to the default NetMHCpan reference data set) is shown in a bar plot. Fitted Gaussian curves belong to the thresholds $\leq 5\%$, $\leq 10\%$ (top binders), and $> 50\%$ (non-binders). We performed the predictions using all viruses of H1N1, H2N2, and H3N2 in the periods 1977-2008, 1957-1968, and 1968-2015.

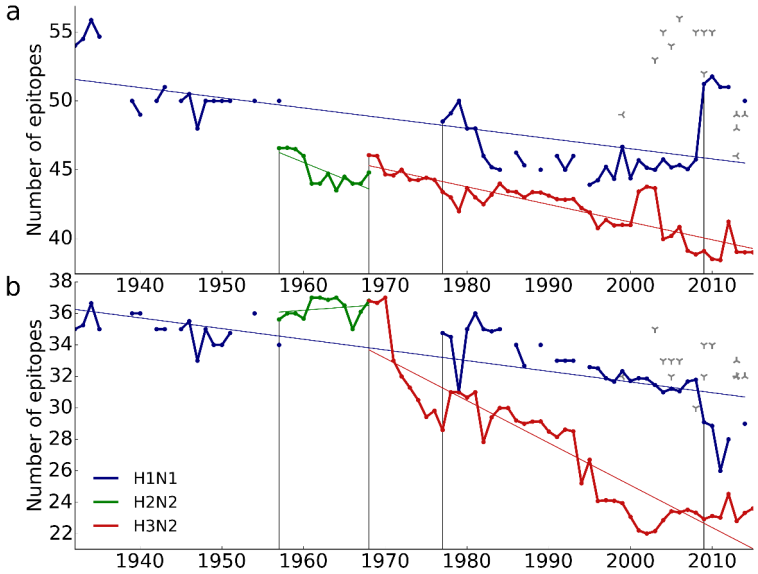


Figure S2.15: Dynamics of CTL epitopes per HLA locus. The number of CTL epitopes per virus in H1N1, H2N2, and H3N2 for HLA-A (**A**) and HLA-B (**B**) alleles. In H1N1, there is a decreasing (non-significant) trend for HLA-A and HLA-B. The year-to-year decrease is most pronounced in H3N2 ($p = 0.045$ and $p = 0.093$ for HLA-A and HLA-B, respectively; Wilcoxon signed-rank test; Methods). Avian viruses are of H5N1 (ㅅ), H7N9 (ㅛ), and H9N2 (ㅓ) subtypes.

Chapter 3

Estimation of age-specific susceptibility to influenza in the Netherlands and its relation to loss of $CD8^+$ T-cell memory

Christiaan H. van Dorp[1,2], Rutger G. Woolthuis[1,2], Jeffrey H. C. Yu[2], Rob J. de Boer[1], and Michiel van Boven[2]

[1]Utrecht University, Utrecht, The Netherlands
[2]National Institute for Public Health and the Environment, Bilthoven, the Netherlands

Submitted

Abstract. The magnitude of influenza epidemics is largely determined by the number of susceptible individuals at the start of the influenza season. Susceptibility, in turn, is influenced by antigenic drift. The evolution of influenza's B-cell epitopes has been charted thoroughly, and only recently evidence for T-cell driven evolution is accumulating. We investigate the relation between susceptibility to influenza, and antigenic drift at $CD8^+$ T-cell epitopes over a 45-year timespan. We estimate age-specific susceptibility with data reported by general practitioners, using a disease-transmission model in a Bayesian framework. We find large variation in susceptibility, both between seasons and age classes. Although it is often assumed that antigenic drift drives the variation in susceptibility, we do not find evidence for a relation between drift and susceptibility in our data. This suggests that other factors determining the variation in susceptibility play a dominating role, or that complex influenza-infection histories obscure any direct effects.

3.1 Introduction

Since the year 1970, the Dutch research institute NIVEL has been using a network of general practitioners (GPs) to estimate the weekly incidence of influenza-like illness (ILI) in the Dutch population [52]. This has resulted in one of the longest ILI-consultation time series available worldwide. The time series contains information about influenza epidemiology, and has enabled analysis of, among other things, the effect of humidity, school holidays [13], waning immunity [241], the impact of vaccination [143], and the validity of models of antigenic evolution [183]. For many of these studies it is presumed that the time series contains implicit information on the immune status of the population.

Immunity to influenza is of humoral (B cells) and cellular (T cells) nature, where the B cells and the antibodies they produce have received most attention. IgA antibodies against epitopes on hemagglutinin (HA) and neuraminidase (NA) provide neutralizing immunity, from which influenza is escaping by means of antigenic drift and shift, as visualized by antigenic cartography [12, 119, 166, 210]. Well-matching vaccines can be efficient in preventing infection [207], but require frequent updates, and vaccine mismatches are not unusual [17, 218]. Cytotoxic T-lymphocyte (CTL) responses against peptidic epitopes of influenza's internal proteins can provide long-term protection [197] because these epitopes are highly conserved. Conserved T-cell epitopes are believed to be the cause of the relatively low illness severity in adults during the 2009 H1N1 pandemic [214], and the 1957 H2N2 pandemic [56], although in both cases antibodies against the more conserved stem of HA could also have played a role. Not surprisingly, these T-cell epitopes are considered to be good targets for a long-lasting, universal vaccine [17, 213].

By cleverly comparing substitution rates within and outside epitope regions in human and swine influenza A virus (IAV), Machkovech *et al.* [140] recently demonstrated positive selection inside CTL epitopes of the nucleoprotein (NP). We also studied the evolution of IAV on the level of CTL epitopes, and found that from 1968 onwards, the number of epitopes in H3N2 has been decreasing [238, Chapter 2]. Both results support the idea that IAV is not only adapting to the human population by escaping from humoral immunity, but also—albeit more slowly—from cellular immunity.

On the clinical level, T-cell mediated immunity is receiving increasingly more attention. Using animal models and human cohort studies, it has become clear that cellular immunity can be responsible for reduced viremia, or even asymptomatic infection [85, 116]. T-cell memory does not provide neutralizing immunity. However, the role of T-cell memory in asymptomatic infection, which in turn reduces infectiousness, leads us to suspect that cellular immunity can have a profound effect on the epidemiology

of IAV. Furthermore, since CTL responses provide long lasting protection [197], the effects of CTL escape can act on a longer time-scale than antibody-antigenic drift.

In this paper, we study the effect of CTL and antibody (Ab) immune escapes of influenza on its epidemiology. We need two types of data for this study: the fraction of susceptible individuals, and the rate of Ab- and CTL-antigenic drift. In order to find the fraction of susceptible individuals, we fit a deterministic ordinary differential equation (ODE) model for transmission to ILI-consultation data. Since this model is quite complex, as the parameters to be estimated can be age- and season-specific, we employ a Bayesian framework, using Markov chain Monte Carlo (MCMC) for parameter estimation. For the other type of data, the antigenic drift, we use the average between-season Ab-antigenic drift as measured by Bedford *et al.* [12]. As a measure of CTL-antigenic drift, we use the disappearance of epitopes, or "epitope loss", as defined recently [238, Chapter 2]. The measure "epitope loss" can roughly be seen as the number of epitopes that disappear from the virus from one year to the next.

Since we expect any relation between susceptibility and antigenic drift to be confounded by age, we stratify the data and model by age class. Incidentally, this allows us to normalize the susceptibilities of the older age classes with the susceptibility of the youngest children, who are likely to be immunologically naive [cf. 56]. In this manner, we correct for external factors shaping epidemic size, that differ between seasons, but have the same effect on the age classes. Such effects include virulence and weather.

Contrary to expectations, we find no significant relation between either measure of antigenic drift, and susceptibility. We discuss explanations for the lack of this relation. This seems to contradict some of the findings presented by Bedford *et al.* [12], who do find some evidence for a relation between antibody-antigenic drift, and the size of the epidemic in a period of two decades in the USA. We find that in the computation of antigenic drift, correct timing of epidemic seasons is essential, and ignoring this may easily lead to false conclusions. Furthermore, we argue that complex individual infection histories may impair our ability to use antigenic drift as a predictor for epidemic hazard.

3.2 Results

We first discuss the details of the data and our model. Then we validate some of our estimates against independent data, and subsequently our estimates are interpreted. Finally, we compare susceptibility with CTL- and Ab-antigenic drift.

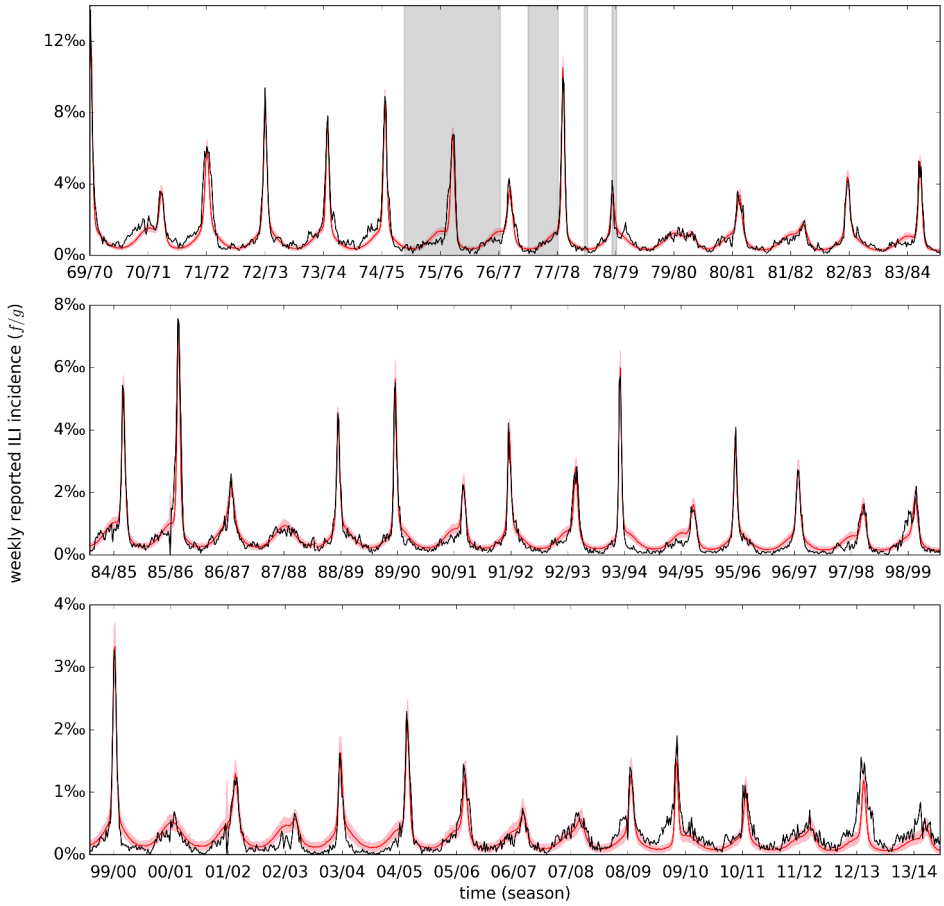


Figure 3.1: ILI data and model fit. Shown is the aggregated weekly ILI incidence (i.e., summed over the age classes), calculated by dividing the weekly number of reported cases (f) by the catchment population size (g) (black line), and the model fit (red line and pink band). The model fit is based on simulations using 200 samples from the posterior distribution of the baseline model. The red line represents the median incidence of the simulated data, and the pink band indicates the 2.5 to 97.5 percentile. The gray background color highlights weeks for which the age stratification and the catchment population size had to be imputed. Notice that the scale on the vertical axis is different for the three panels, as the reported incidence is declining.

3.2.1 Data and model selection

An overview of the (aggregated) ILI data and the fitted model is given in Figure 3.1. From the first season (1969/70) to the last (2013/14), the fraction of reported ILI is gradually decreasing. This holds for both the epidemic peaks in the winters, and for the lower off-season weeks. There is a variety of potential explanations for this trend [47]. First, the circulation of ILI causing pathogens could be slowly decreasing. For the elderly, this could be partially due to changes in vaccination policy [143]. Second, it could be that people have become healthier, or the virus less virulent [62]. Finally, it could be that the reporting tendency has decreased over the last 45 years, because people less frequently visit their GP when experiencing mild symptoms. The latter explanation is most simple, seems biologically plausible, and is incorporated in the current model. The probability that a person with ILI consults their GP is denoted q, and we model the downward trend in reporting by allowing the probability q to depend on time (see Methods).

An ILI data point consists of the number of reported ILI cases (f), and the catchment population size during that week (g, see Figure S3.2). The catchment population is a small sample from the entire Dutch population, and we assume that the total population is large enough so that the influenza epidemic can be modeled with a deterministic model of ODEs. ILI can be caused by a range of pathogens other than influenza (A and B), such as respiratory syncytial virus, rhinovirus, parainfluenza, and *Mycoplasma pneumoniae*. We refer to ILI that is *not* caused by influenza as background ILI.

The characteristic epidemic peaks are mostly caused by influenza, and are modeled with a transmission model (see below). Concerning the background ILI, a clear seasonal effect is visible (Figure 3.1), even during the off-season weeks. For this reason, we use a descriptive model that takes seasonality into account. This approach is similar to the one taken by Van Noort *et al.* [167]. Furthermore, McDonald *et al.* [143, 144] used weekly laboratory surveillance data to discriminate between background ILI and influenza. The ILI in the shoulder of the epidemics turns out to be mainly caused by pathogens other than influenza.

Throughout, we stratify the ILI data into 6 age classes (Methods). The contacts between these groups are modeled using an age-specific contact structure that was derived with data from a human contact study (Mossong *et al.* [158]; see Methods). The ODE model for influenza therefore consists of 4×6 compartments. For each age class, we have susceptible individuals (S), early and late stage infected individuals (I_1, I_2), and recovered individuals (R; not explicitly modeled). The two infection stages ensure a more realistic (Erlang) distribution of the length of the infectious period, and we assume no difference in infectiousness between the stages. The mean duration of the infection ($1/\gamma$) is not estimated; rather, we take $1/\gamma = 3.0$ days, which is within the plausible range [32].

scenario	relations			k	ΔWAIC	ΔWBIC
baseline	r_1			349	-	-
trended susceptibility	-			355	0.4	10.4
single prior for susceptibility	r_1	r_3		339	8.9	304
susceptibility independent of age	r_1	r_3	r_4	114	7285	5820
constant background ILI	r_1 r_2			347	21 642	21 356

Table 3.1: Model selection. Models are compared with WAIC and WBIC (see Methods). The number of parameters used in each model is denoted by k. Each one of the models is defined by relations between the parameters (see Methods), defined by: $r_1 = \{\alpha_{\text{susc},1} = 0, \ldots, \alpha_{\text{susc},6} = 0\}$, $r_2 = \{b_{\text{amp}} = b_{\text{phase}} = 0\}$, $r_3 = \{\mu_{\text{susc},1} = \cdots = \mu_{\text{susc},6}, \alpha_{\text{susc},1} = \cdots = \alpha_{\text{susc},6}, \sigma^2_{\text{susc},1} = \cdots = \sigma^2_{\text{susc},6}\}$, $r_4 = \{S_{0,1} = \cdots = S_{0,6}\}$. The best model (baseline) assumes relation r_1, meaning that S_0 is stationary. The baseline model has WAIC = 95 926 and WBIC = 97 073. For the model with constant background ILI we assume relation r_2 in addition to r_1, meaning that the background ILI is the same throughout the year. The model that only incorporates a single prior distribution for susceptibility (i.e. no age effect) assumes relation r_3 in addition to r_1. The model with age-independent susceptibility incorporates relation r_4 in addition to r_1, and (trivially) r_3.

For the initial conditions of the model, we need the fractions of susceptible individuals (S_0) at the beginning of the seasons. In our model the parameter S_0 is the probability that an uninfected individual is infected upon contact with an infectious person, and becomes sufficiently ill to have a probability q to visit a GP. Individuals that do become infected, but experience mild disease (e.g. due to T-cell memory [56, 214]), are therefore not considered to be susceptible, and are assumed to be hardly infectious, due to limited viral shedding [120]. Notice that there is a slight mismatch between our interpretation of S_0, and the usage of this parameter in the model. This discrepancy comes from the fact that it is hard to distinguish susceptibility from infectiousness in ODE models. For our purpose, both susceptibility and infectiousness are of interest. Hence, the current parameterization is suitable for our goal. This issue is discussed in more detail in the Supplementary material.

We use a mixed effects model for the susceptibility parameter S_0. That is, *a priori*, the S_0 are assumed to be sampled from Logit-Normal distributions, with unknown mean (μ_{susc}) and standard deviation (σ_{susc}). We consider versions of the model where μ_{susc} and σ_{susc} are dependent on the age class, and also a version where μ_{susc} is a linear function of the season s (see Estimation of the parameters).

Several models are compared with information criteria WAIC and WBIC (Table 3.1 and see Methods). Our analyses show that age-dependent susceptibility and periodic background ILI are essential elements of the model. The model that incorporates an age effect for S_0 is favored over a model where every S_0 has the same prior distribution. The

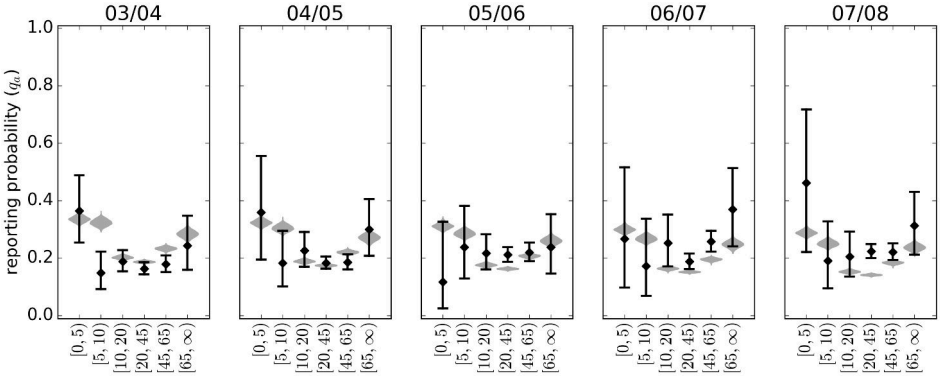


Figure 3.2: Reporting estimated from ILI data compares well to GIS reporting frequencies. The gray violin plots show the marginal posterior distribution of the reporting probabilities (q), which are used to scale reported ILI to influenza and background incidence. The black diamonds show the reporting frequency from the GIS data obtained from 2003 to 2008. The black whiskers indicate 95% confidence intervals (obtained using Jeffrey's method).

model incorporating a time effect for S_0 fits the data well, but has too many (effective) parameters.

3.2.2 Estimated reporting matches independent observations

In our analyses the reporting probability (q) is mainly a nuisance parameter that ties the observed number of ILI cases to the underlying circulation of influenza and background ILI. It is important, however, that estimates of the reporting do not systematically skew parameters that are truly of interest in this study. To provide external validation of our reporting rate estimation, we have used data from an independent study on influenza incidence [GIS, 67, and see Methods]. Reassuringly, our estimates and those from GIS correspond well (Figure 3.2). Even the U-shape of the age-stratified reporting is captured by our estimates. Naturally, we can not use the results from this small number of consecutive years to extrapolate the validity of our reporting estimates to the entire time series (Figure S3.1), but the correspondence between the estimate of q and this independent data demonstrates our ability to extract this information from the ILI data.

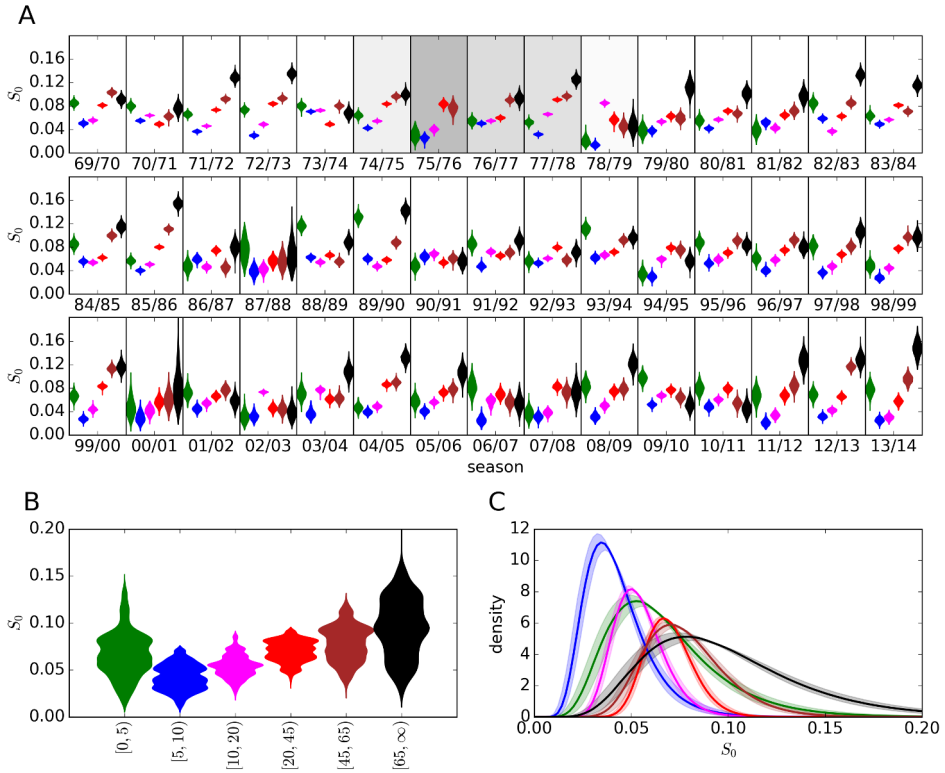


Figure 3.3: Susceptibility estimates stratified by season and age class. A. The violin plots show the marginal distribution of the susceptibility parameters $S_{0,a}$ for $a = 1,\ldots,6$. The color coding for age classes is as follows: green: $[0,5)$ years, blue: $[5,10)$ years, cyan: $[10,20)$ years, red: $[20,45)$ years, brown: $[45,65)$ years, black: $[65,\infty)$ years. The gray background indicates missing age stratification during a season, with brightness proportional to the number of non-missing weeks (cf. Figure 3.1). **B.** The samples from the posterior distribution are aggregated by age class. **C.** Posterior predictive distributions of age-specific susceptibilities, i.e. the Logit-Normal distributions of the parameters $S_{0,a}$ are shown ($a = 1,\ldots,6$), with hyper parameters taken from the posterior distribution. The lines indicate the median densities, and the bands represent the interquartile range.

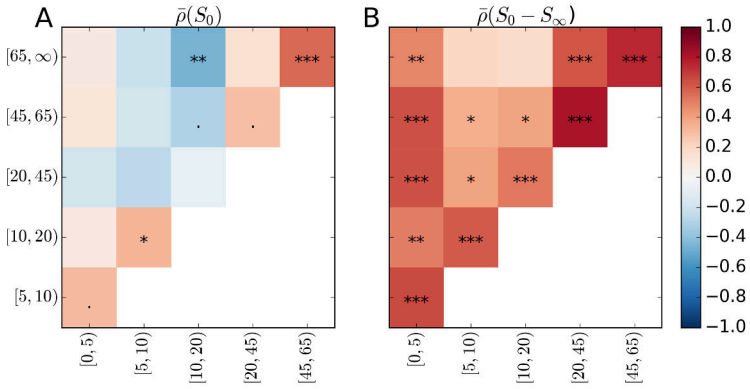


Figure 3.4: Correlations between estimated age-specific susceptibilities. A. Every square is colored according to the expectation of the Spearman rank correlation between S_{0,a_1} and S_{0,a_2} (with $1 \leq a_1 < a_2 \leq 6$). For instance, in seasons where individuals in the age class $[45, 65)$ are highly susceptible, the elderly (age class $[65, \infty)$) tend to be susceptible as well. The stars indicate the expectation of the significance level: '·': $0.1 \geq p > 0.05$, '∗': $0.05 \geq p > 0.01$, '∗∗': $0.01 \geq p > 0.001$, '∗∗∗': $0.001 \geq p$. The seasons 1975/76 and 1978/79 are not included in the analysis, because of the missing age-stratification at the time of the epidemic peak. **B.** Similar to (**A**), but then comparing the cumulative incidences of the age classes, i.e. the fraction of individuals that has been infected during the season ($S_0 - S_\infty$, where S_∞ denotes the fraction remaining susceptible at the end of the epidemic).

3.2.3 Susceptibility differs between and within age classes

The main outcome of the fitted Bayesian model is a set of estimates of the susceptibility to influenza of individuals in the 6 age classes at the beginning of 45 influenza seasons. Figure 3.3 shows an overview of these 6×45 estimates. Clear differences exist between the age classes (Figure 3.3B). The highest susceptibilities can be found among the elderly ($[65, \infty)$ years), but this age class also shows the most variation and uncertainty. The elderly are followed by the youngest age class ($[0, 5)$ years) and the age class $[45, 65)$ years. Children in the age class $[5, 10)$ years are least susceptible. On the other hand, age is not a perfect predictor for susceptibility, as the posterior predictive densities of the susceptibilities show a large overlap (Figure 3.3C). Summarizing, susceptibility is determined by both age and calendar time (epidemic season). The cause of the effect of age on susceptibility must be sought in individual health states and immunological history, and that of the epidemic season in antigenic drift, viral fitness, and environmental factors.

We observe considerable variation of susceptibility both within and between age classes. As we will argue now, this variation is not entirely random. By performing pair-

wise comparisons between age classes over all seasons (Figure 3.4A), we make the following observations: (i) The susceptibility of closely related age classes is correlated significantly, or has a tendency to be correlated, except for the age classes $[10,20)$ and $[20,45)$ years. (ii) Susceptibility of children in the age class $[0,5)$ years is only somewhat correlated with age class $[5,10)$ years. (iii) Susceptibility of adults and elderly is not, or negatively correlated with the susceptibility of the age classes $[5,10)$ and $[10,20)$ years.

These observations are understandable in light of immunological history. Individuals in closely related age classes are more likely to share a common infection history than individuals with a greater age difference, which would explain observation (i). Since children in the age class $[0,5)$ years are more likely to be immunologically naive, their susceptibility does not resemble that of older individuals, although a certain overlap is again to be expected with the age class $[5,10)$ years, hence observation (ii). Observation (iii) is more difficult to explain, and might be related to "original antigenic sin" (OAS), or the related "antigenic seniority" [125]. Recently, it was shown that antigenic seniority is likely to generate negative correlations between young and old birth cohorts with respect to susceptibility to severe avian influenza infection [81]. Such a mechanism could work more generally. If an influenza antigen resembles another recent antigen, then individuals in the age class $[10,20)$ years mount an effective immune response, and their susceptibility is low. Although older individuals also encountered the recent antigen, they may have failed to create memory against it (due to OAS), and suffer the consequence: increased susceptibility for currently circulating viruses. Vice versa, when older individuals respond well, this may be due to similarity to an old antigen (cf. the antigenic thrift model [235]), which individuals in the age class $[10,20)$ years never encountered. Again, the fact that we find intuitive correlations between the estimated susceptibilities of the various age classes illustrates that we are obtaining meaningful estimates of S_0 from the MCMC procedure.

An alternative measure for the immune status of the population is the attack rate ($S_0 - S_\infty$, see Figure 3.4B). Arguably, it is more straight-forward to extract attack rates from ILI data than susceptibility. In fact, the attack rate (or average incidence) has been used in similar studies before [12]. In the standard SIR model, the relation between susceptibility and attack rate is quite simple, albeit non-linear. The non-linearity can be described as herd immunity. Herd immunity becomes more important when multiple compartments (in our case age classes) are involved; the attack rate within a class becomes a poorer predictor for the susceptibility of this class, since the other classes interfere. This can be seen in Figure 3.4B. In a way, estimating susceptibility with a compartmental model could be seen as a method to de-correlate age-stratified incidence data.

Using the estimates of S_0 and the contact matrix, we can compute for each season the effective reproduction number $\mathcal{R}_{\text{eff}}$ at the start of the epidemic (Figure 3.5), still as-

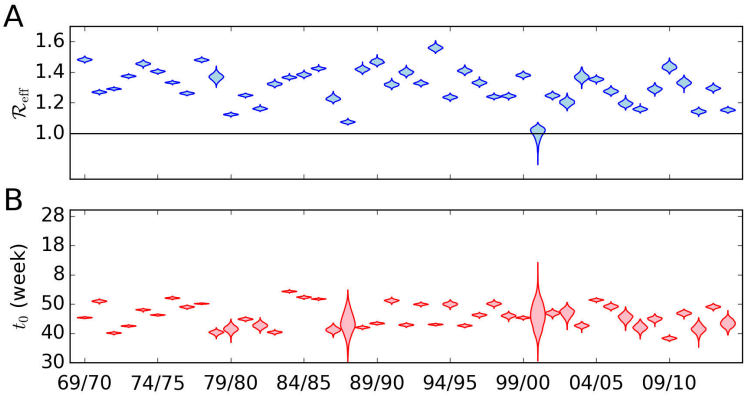


Figure 3.5: Estimates of the reproduction number and the onset of the epidemics. A. The posterior densities of the compound parameter $\mathcal{R}_{\text{eff}}$ (the effective reproduction number) for the 45 seasons. The horizontal black line indicates the epidemic threshold $\mathcal{R}_{\text{eff}} = 1$. **B.** Estimates of t_0, the start of the 45 epidemics (week number). The precision $(\text{sd}(t_0)^{-1})$ of the start of the epidemic is strongly correlated with the posterior mean of $\mathcal{R}_{\text{eff}}$ (Spearman $\rho = 0.85$, $p < 10^{-12}$).

suming that individuals that experience mild disease hardly contribute to the epidemic. The effective reproduction number is related to the basic reproduction number in the sense that they are equal when there is no pre-existing immunity. On average, $\mathcal{R}_{\text{eff}}$ equals 1.31 (95% CrI: $[1.07, 1.49]$; aggregated over all seasons). It is to be expected that the epidemic in a season with a high $\mathcal{R}_{\text{eff}}$ starts earlier than epidemics in seasons with low $\mathcal{R}_{\text{eff}}$, but we do not find any evidence for this in the data. We do find a strong correlation between the $\mathcal{R}_{\text{eff}}$ of a season, and the precision with which we can estimate the start of the epidemic (Figure 3.5), as the interference of background ILI is stronger for small epidemics. Perhaps our inability to find a relation between timing and reproduction number is due to the lack of precision with which we can estimate this timing.

Some of the estimated reproduction numbers can easily be related to the circulation of specific strains in the Netherlands. For instance, the relatively high reproduction number at the start of the 1993/94 epidemic season can probably be attributed to a strain from the BE92 (H3N2) cluster, which in antigenic space is far away from the previous BE89 cluster [210]. In addition, the cluster transition around the year 1993 coincides with a large CTL-antigenic jump [238, Chapter 2]. Similarly, the high $\mathcal{R}_{\text{eff}}$ in 2009/10 is due to the pandemic H1N1 strain. The reappearance of H1N1 in 1977 (in the form of the A/USSR/90/77 strain) might be responsible for the high $\mathcal{R}_{\text{eff}}$ in 1977/78. For 1969/70, we only have data for the second half of the epidemic. The fact that we are able to estimate susceptibility and a reproduction number for this season,

is due to the mixed effects model for S_0. The high $\mathscr{R}_{\text{eff}}$ for this season corresponds well with estimates from England and Wales [62], and might be due to the re-assortment event that led to the replacement of H2N2 by H3N2. No H3N2 strains circulated during the 2000/01 season [47], resulting in an estimated $\mathscr{R}_{\text{eff}}$ around 1.

3.2.4 CTL- and Ab-antigenic drift do not predict (relative) susceptibility

Previously, we calculated how many CTL epitopes appear or disappear from IAV between consecutive seasons [238, and see Methods]. Since the escape of an epitope should lead to loss of CTL memory in some individuals, we test whether the number of lost epitopes in H3N2 can explain susceptibility as estimated above. We focus on the H3N2 subtype, since this subtype circulates the most, causes the most severe disease, and evolves most rapidly. It is expected that an effect of CTL epitope loss on estimated susceptibility is small in the youngest children, because most of them have not been infected before, and tend to lack CTL memory against influenza. Additionally, the fact that CTL-epitope regions in IAV are relatively conserved [140] could be an indication of a fitness cost associated with mutations in these regions [16]. A strain that has many mutations in these epitopes, could therefore suffer from diminished infectiousness, when compared to the wild-type form in immunologically naive populations.

Apart from intrinsic fitness effects, variation of the susceptibility of the youngest children should also reflect environmental influences on influenza transmission, such as (indoor) humidity and temperature [133, 134]. We normalize for such yearly effects by considering the susceptibility of the age classes, relative to the susceptibility of the youngest age class (henceforth "relative susceptibility"). In the following analyses, we leave out seasons dominated by H1N1 (1983/84, 1986/87, 2000/01, 2007/08, 2009/10, 2010/11). Likewise, seasons with missing age-stratified data around the epidemic peak we also removed from the analyses (1975/76, 1978/79).

We use the antigenic distance between strains [12] as a measure of Ab-antigenic drift. In short, this antigenic distance is derived from hemagglutinin inhibition (HI) assays using Bayesian multi-dimensional scaling (BMDS). The location of the strains along the first antigenic dimension is informed by a "drift prior", in order to solve identifiability issues. Displacement along this axis is therefore a measure of antigenic drift. The antigenic drift between epidemic seasons is derived by first averaging the first coordinates of all strains in a particular year, and then taking the difference of the averages. No correlation between Ab-antigenic drift and relative susceptibility can be found (Figure 3.6A and Table 3.2).

We use "epitope loss", as defined previously [238, and see Methods], as a measure for CTL-antigenic drift. Epitope loss measures the number of epitopes that were present

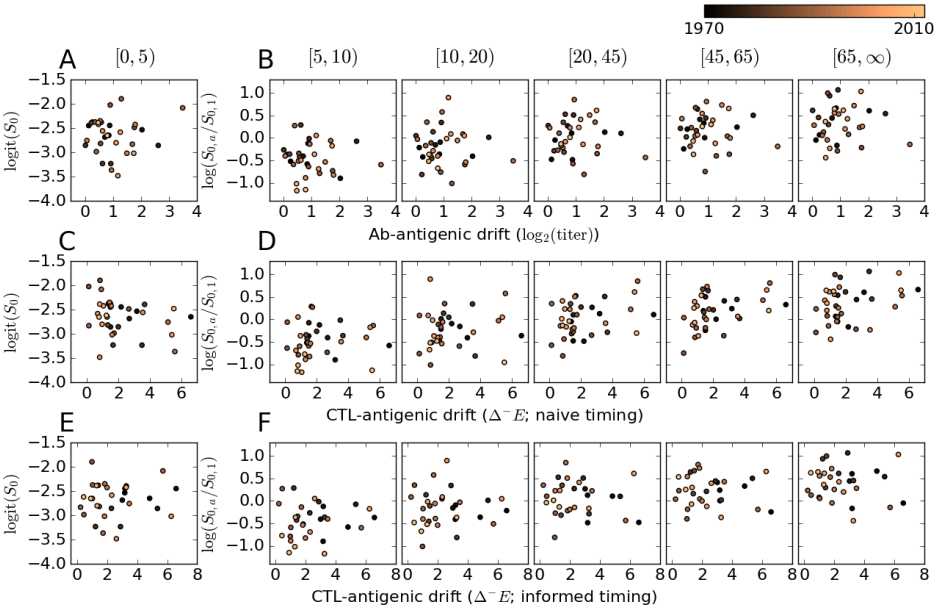


Figure 3.6: Susceptibility plotted against antigenic drift. Each dot corresponds to an epidemic season (the tone indicates calendar time). For the age class $[0,5)$, the posterior mean absolute susceptibility is used (**A**, **C**, **E**), and for the other age classes, the posterior mean susceptibility relative to the first age class (**B**, **D**, **F**). We test different antigenic drift values for their predictive power. (**A**, **B**) Antibody-antigenic drift, (**C**, **D**) CTL-antigenic drift with "naive" timing (i.e. isolates are assigned to a season based on only their calendar year), (**E**, **F**) CTL-antigenic drift with "informed" timing (i.e. periods with significant IAV circulation are derived from the fitted model, and missing isolation dates are sampled). The Spearman correlation coefficients are listed in Table 3.2.

age class	Ab-Antigenic drift		CTL-Antigenic drift			
			naive		informed	
	ρ	p	ρ	p	ρ	p
$[0,5)$	−0.15	0.40	−0.29	0.09	0.03	0.88
$[5,10)$	−0.10	0.59	0.25	0.14	0.09	0.62
$[10,20)$	0.02	0.90	0.16	0.36	0.05	0.78
$[20,45)$	0.09	0.62	0.24	0.17	−0.10	0.59
$[45,65)$	0.06	0.73	0.37	**0.02**	−0.08	0.67
$[65,\infty)$	0.12	0.50	0.29	0.08	−0.10	0.61

Table 3.2: Correlations between antigenic drift and (relative) susceptibility. The correlation is calculated between the posterior mean (relative) susceptibility, and measures of antigenic drift. For the age class $[0,5)$, the absolute susceptibility is used. Listed are the Spearman correlation coefficients (ρ), and the corresponding p-value. The susceptibility and antigenic drift values are plotted in Figure 3.6.

in one year, but absent (escaped) in the next. Intriguingly, we do find a significant positive correlation between CTL-antigenic drift and relative susceptibility, but only in the age class $[45,65)$ (Figure 3.6B and Table 3.2). Positive trends can be observed in some of the other age classes, and a negative trend between CTL-antigenic drift and the absolute susceptibility of the youngest age class.

The between-season Ab- and CTL-antigenic drift used in the above analyses has been computed using a number of simplifying assumptions, including the way isolates are assigned to epidemic seasons (i), and what isolates are selected for the analyses (ii). Assumption (i) can be problematic for two reasons. Firstly, the older strains are often only dated at the resolution of calendar years. In the Northern hemisphere, influenza epidemics consistently occur during the winter months, and hence for these strains it is difficult to determine which season they should be assigned to. Secondly, the timing of the epidemic season differs from season to season (Figure 3.5), in such a way that for some seasons (e.g. 1978/79), the epidemic falls mostly in the first calendar year (1978), while for other seasons (e.g. 1990/91), the epidemic falls in the next calendar year (1991). Assumption (ii) can be problematic when isolates are used that circulate during the summer. Many of them are sampled in the Southern hemisphere, and it is not clear if they can be used as representative strains during the winter epidemic, or for which of the two possible winter epidemics (cf. assumption (i)). Furthermore, sparse sampling of isolates can lead to large uncertainty in antigenic drift.

We resolve both problematic assumptions using a bootstrapping procedure. Using our epidemiological model, we estimate the periods that most likely contains the influenza

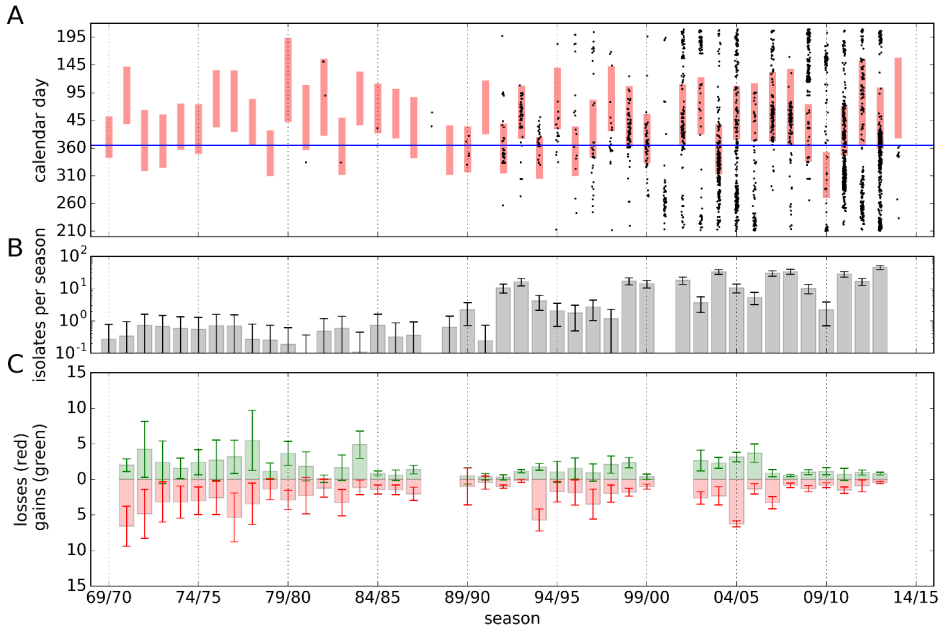


Figure 3.7: Estimation of CTL-antigenic drift. A. Time intervals of influenza circulation in the Netherlands, and known sampling dates of the isolates. The red bars indicate the epidemic periods, and the black dots the sampling dates of the isolates. **B.** The yearly number of isolates used for the analyses. The whiskers indicate the standard deviation due to the bootstrapping procedure (see Methods). **C.** The epitope gain and loss, as defined by Equation (3.1). The whiskers indicate the standard deviation, estimated by bootstrapping.

epidemic (Figure 3.7A; pink bands). For isolates that have interval censored isolation dates, we sample random dates from a distribution estimated from the known isolation dates (Figure S3.3). We then sample $N = 1000$ isolates from the total set of viruses [see 238, Chapter 2] to address assumption (ii), and assign an isolate to a season if it falls into the estimated time interval. Isolates that do not fall into an epidemic season are discarded. This resolves assumption (ii). Using a sampled set of isolates and isolation dates, we re-compute the epitope loss (and gain; see Methods). This procedure is repeated a 1000 times. The result of this more accurately computed epitope loss (and gain) is shown in Figure 3.7C. Unfortunately, when the timing-aware CTL-antigenic drift replaces the previously used timing-naive CTL-antigenic drift, the previous correlation with relative susceptibility in the age class $[45, 65)$ years is lost (Figure 3.6F; Table 3.2).

3.3 Discussion

Using long-term data of ILI reported to GPs, we estimated the yearly age-specific susceptibility of the Dutch population to influenza. Our estimates were obtained using a transmission model in a Bayesian framework, so that all parameters have a clear-cut biological interpretation. We found large variation in susceptibility within age classes, and interesting correlations between age classes that can be explained by expected differences in immunological memory. In contrary to expectation, we were unable to find a relation between Ab- and CTL-antigenic drift in the H3N2 subtype and relative susceptibility.

A large body of work exists where parameters are estimated from ILI (or related) time series. For the Dutch ILI data, this was done by e.g. Te Beest *et al.* [13], Ratmann *et al.* [183], Xia *et al.* [241]. Related studies using non-Dutch data were presented by Baguelin *et al.* [8], Goeyvaerts *et al.* [76], Yang *et al.* [242]. Most notably, Ratmann *et al.* [183] use NIVEL's ILI data in synergy with influenza HA sequence data to test evolutionary hypotheses about the virus. Specifically, using approximate Bayesian computation, they fitted the epochal evolution model of Bedford *et al.* [11] to the ILI data.

The fact that it seems so difficult to find relations between antigenic drift and susceptibility is curious, and has been noticed before [34, 167]. Yet, the belief that variation in attack rates is caused by antigenic drift is widespread. Below, we discuss several reasons for not finding any relation between antigenic drift and susceptibility in detail, but it should be noted that our result, and lack of other epidemiological evidence indicates that the relation between antigenic drift and susceptibility is not as strong and direct as often assumed.

Although we are comparing estimated susceptibility to IAV with disappearance of CTL epitopes, we do not assume that CTL-memory provides neutralizing immunity against IAV infection. More likely, any effect of CTL-memory should be sought in reduced viremia, or enhanced clearance [77], and the increased likelihood of asymptomatic infection [85]. If asymptomatic infection coincides with decreased transmission, asymptomatically infected individuals can be counted as non-susceptible, since a model with an asymptomatic compartment is equivalent to the usual SIR model [167, and Supplementary material].

We compared relative susceptibility with a simple measure of antigenic drift. In reality each individual will have its particular immunity against future influenza strains, due to the many influenza infection histories that are possible [64], and therefore each individual experiences the antigenic drift or shift differently [36, 37]. As an extreme example, HA-imprinting and the replacement of H1 with H2 in 1957 and then H3 in 1968 is likely the reason why birth year explains susceptibility to either severe H7N1, or severe H5N1 avian influenza infections [81]. Because we consider T-cell responses, HLA polymorphism complicates the matter even further; each HLA haplotype results in a different set of possible epitopes. Possibly, an agent-based model can be used to generate individual infection histories, giving a more complete measure of the effects of CTL-antigenic drift, that can then be compared with susceptibilities measured from ILI data. However, such an approach is not straightforward. For instance, as mentioned by Ratmann *et al.* [183], estimation of the effect of immune histories on susceptibility is sensitive to small uncertainties in the historical sequences and ILI data. Our previous considerations about epitope loss indicate that these uncertainties may not even be that small.

Another source of uncertainty is the fact that ILI is not a very specific predictor for influenza. The positive predictive value (PPV) of ILI has been estimated at $\approx 50\%$ for influenza during the peak ILI weeks of the seasons 2003/04 to 2011/12. We attempt to correct for this relatively low PPV by using a phenomenological model for the background ILI. When instead the background ILI is assumed to occur at a constant rate throughout the year (the "constant background ILI" scenario; Table 3.1), the estimates of $\mathscr{R}_{\text{eff}}$ tend to stay away from the epidemic threshold ($\min \mathscr{R}_{\text{eff}} = 1.11$), since background ILI wrongfully has to be explained by influenza. However, we expect some model misspecification by ignoring the fact that background ILI, akin to influenza, is caused by infectious pathogens.

Our assumption about the mechanism underlying the trend in the reporting rate is consistent with other studies [47], but not based on hard evidence. Although our estimates for more recent years are consistent with an independent data source (GIS), the high reporting rate during the seventies could be an artifact of our model. Van Noort *et al.* [167] have hypothesized that not infectiousness, but reporting rate is influenced by weather, and use this to explain variation in reported incidence. However, such

a hypothesis is not sufficient to explain the strong decrease in reported ILI during the last five decades. In a later study [168], differences in reporting rate between countries are attributed to cultural differences. The differences in reporting rate between countries are of the same order of magnitude as the difference within the Netherlands between 1970 and 2014. Another possibility is formulated by Morens *et al.* [156], who hypothesize that for evolutionary reasons, the virulence of influenza could be decreasing, while infectiousness is retained; a host that is not put to bed will spread a disease more efficiently. Although intriguing, this idea does not fit well with our model, since not only influenza reporting decreases, but also reporting of background ILI. Finally, the fitness of seasonal influenza could be slowly deteriorating, possibly due to continuous antigenic drift needed for immune evasion, as argued by Fleming and Elliot [62] based on an ILI time series from England and Wales. However, following the same reasoning as above, most ILI-causing pathogens should then be subject to fitness erosion. This includes (in hindsight) the pandemic H1N1 virus from 2009/10, since the incidence during this season is not strikingly different from its surrounding years (Figure 3.1).

Some of the decreasing ILI in the age class $[65, \infty)$ could be the result of vaccination policy [143] and uptake [212]. As of the year 1997, vaccination against influenza for individuals in this age class is free of charge in the Netherlands. Despite the fact that we use a trended q to account for the decreasing ILI, one might still expect a difference in susceptibility before and after 1997 for the elderly. However, we cannot find evidence of an effect of vaccination by making this comparison ($p = 0.47$, t-test based on posterior means). Analogous to antigenic drift and susceptibility, the difficulty with finding associations between vaccine uptake and ILI incidence may be due to complex infection histories itself [212],

Recapitulating, our analyses have shown that, on the one hand, there are systematic and substantial differences between age classes in their susceptibility to IAV infection, with 5 to 20 year-olds being least susceptible, and infants and young children ($[0, 5)$ years) and older adults (65 years and older) being most susceptible. On the other hand, there is also substantial variation in susceptibility between years, thereby precluding attempts to quantitatively predict age-specific susceptibility with meaningful precision. Although it is generally believed that these patterns of differences in susceptibility between years and between age-groups are molded by viral evolution and pre-existing immunity, we were unable to find evidence for this hypothesis in our data. This was true both for humoral immunity mediated by antibodies directed against the hemagglutinin protein and also for cellular immunity mediated be cytotoxic T cell responses. Hence, we are led to the perhaps somewhat sobering viewpoint that the complex interplay of viral evolution and pre-existing immunity is highly sensitive to details of the infection histories in narrow age strata [81, 183], and may be fundamentally unpredictable.

3.4 Methods

3.4.1 Data

We made use of a number of data sources. For self-consistence, we here briefly discuss these sources, and our particular usage of the data.

NIVEL. The catchment population covers approximately 1% of the Dutch population, equally distributed over different regions in the Netherlands [47, 52]. Using the ISO week date system, we translated years and week numbers into the number of days since January 1, 1970. This solves the issue of 'leap weeks', which should be treated with care, since they tend to coincide with the epidemic peaks.

The data is stratified into 6 age classes: $[0,5)$, $[5,10)$, $[10,20)$, $[20,45)$, $[45,65)$ and $[65,\infty)$ years, usually indexed by $a = 1,\ldots,6$, respectively. As of 1986, the resolution of the age stratification has been increased. Our age classes are chosen to be compatible with early and more recent ILI data stratification, and to fit well with contact-intensive clusters (not shown) that can be observed in the contact matrix (see below). Some of the ILI data is missing, but has been partially recovered (see Missing data).

Dutch law allows the use of electronic health records for research purposes under certain conditions. According to this legislation, neither obtaining informed consent from patients nor approval by a medical ethics committee is obligatory for this type of observational studies containing no directly identifiable data (Dutch Civil Law, Article 7:458). This study has been approved by the applicable governance bodies of NIVEL Primary Care Database under nr NZR00316.056.

POLYMOD. The POLYMOD study is a prospective survey of social contact patterns in multiple countries [158, 226]. The key assumption for epidemiological studies of respiratory-spread infectious diseases is that conversational contact can be used as a proxy for infectious contact [the social contact hypothesis, 226]. Our contact matrix is based on only the Dutch data, and estimated using a method developed by Van de Kassteele *et al.* [104]. In the study, *participants* (see Figure S3.4) were asked to keep a diary on the individuals with whom they had *contact* during one day. The average numbers of contacts applied to our age classes are given in Figure S3.4.

GIS. The Great Influenza Survey is the Dutch branch of the European Influenzanet project [168], a monitoring system for ILI, using voluntary cohorts approached via the Internet. In the on-line questionnaire, the participants are asked, among other things,

if they had influenza-like symptoms such as fever and coughing, and if they consulted a GP while they had these symptoms.

Most of the GIS data used in this paper is published by Friesema *et al.* [67]. Unfortunately, the age stratification in this report has a slight mismatch with our choice of age classes, and in order to compute confidence intervals, the population sizes of self-diagnosed ILI patients, and the sub-population seeking a GP consult are needed. I. Friesema kindly provided the original data, which enabled us to compute the population sizes for our age stratification.

Epitopes. The full procedure for collecting epitopes is described in [238, Chapter 2]. In short, we collected all complete, human IAV proteomes without ambiguous amino-acids from the GISAID EpiFlu database (`www.gisaid.org`). All known CTL epitopes were downloaded from IEDB (`www.iedb.org`). Epitope sequences that contained shorter epitopes were filtered out. For each epitope e and each calendar year y, the fraction of IAVs that contained the epitope is denoted by $\varphi_{y,e}$. Epitope gain ($\Delta_y^+ E$) and loss ($\Delta_y^- E$) were defined as

$$\Delta_y^{\pm} E = \sum_e \max\left\{0, \pm(\varphi_{y,e} - \varphi_{y-1,e})\right\} . \tag{3.1}$$

In order to estimate the uncertainty of the gain and loss, we applied the following bootstrap procedure. First, we sample (with replacement) a 1000 strains from the total number of 3321 isolates. Using the sample, we compute gains and losses (Equation 3.1) for every year y, whenever the sample contains isolates from both years in the pair $(y-1, y)$. We repeat the sampling and computation $n = 1000$ times. Notice that we produce less bootstrap samples for a year y, whenever few strains were isolated in the year y or the previous year $y-1$.

In order to assign isolates to seasons in an informed manner, we first have to estimate when the epidemics took place. We assign a a week to the epidemic, when the weekly incidence of the total catchment population exceeds 10 cases per 100 000 individuals. Hence, when $\sum_a \Delta S_{a,w} g_{a,w} > 1 \cdot 10^{-4} \sum_a g_{a,w}$ (using notation introduced below). We then define t_{start} as the first day of the first epidemic week and t_{end} as the last day of the last epidemic week. Notice that t_{start} and t_{end} need not be defined; when $\mathscr{R}_{\text{eff}}$ is close to 1, the incidence may never reach the threshold. This leads to epidemic periods that are on average 12.3 weeks long (IQR: $[11.0, 14.0]$ weeks). On purpose, our the epidemic threshold is 5 times smaller than the guideline used by NIVEL (51 cases per 100 000), resulting in slightly longer epidemics. Isolates are then assigned to season s, if they were sampled within the epidemic period $[t_{\text{start}}, t_{\text{end}}]$ of season s. Isolates that do not fall into any epidemic period are ignored. Missing isolation dates are sampled from a distribution inferred from the known dates (Figure S3.3). This distribution is

estimated by fitting a Circular-Gaussian kernel with a bandwidth of 30 days to the set of known calendar days. The bootstrapping procedure is similar as in the naive case. A 1000 strains are sampled (with replacement), then, sampled isolates with missing dates are randomly generated, and finally we assign seasons to the isolates. Epitope gain and loss is computed as usual.

3.4.2 Dynamical model

We model the influenza epidemics using a system of ordinary differential equations (ODEs). The population is partitioned into classes depending on disease status: susceptible (S), infectious (I) and recovered (R). We use an age-stratified SIIR model, with fixed contact rates C (Figure S3.4). The model is given by the following equations:

$$\begin{aligned}
\tfrac{\mathrm{d}}{\mathrm{d}t}\vec{S} &= -\mathrm{diag}(\vec{S})C(\vec{I}_1+\vec{I}_2)\\
\tfrac{\mathrm{d}}{\mathrm{d}t}\vec{I}_1 &= \mathrm{diag}(\vec{S})C(\vec{I}_1+\vec{I}_2)-2\gamma\vec{I}_1\\
\tfrac{\mathrm{d}}{\mathrm{d}t}\vec{I}_2 &= 2\gamma\vec{I}_1-2\gamma\vec{I}_2\,,
\end{aligned} \tag{3.2}$$

where $\mathrm{diag}(\vec{S})$ is the diagonal matrix with the vector $\vec{S}$ on the diagonal. The two infectious phases are used to better model (non exponential) distribution of the length of the infectious period. The initial condition at time $t = t_0$, is given by a small perturbation (of size ε, with $0 < \varepsilon = 10^{-6} \ll 1$) of the disease-free steady state: $\vec{S}_0 \oplus \vec{0} \oplus \vec{0} + \varepsilon v$, where $\vec{0}$ is the zero vector in $\mathbb{R}^6$, and $\oplus$ denotes concatenation of vectors. Here, v is the dominant eigenvector of the Jacobian of system (3.2) at the state $\vec{S}_0 \oplus \vec{0} \oplus \vec{0}$. The vector v has positive I-coordinates and satisfies $\|v\|_2 = 1$. Using the S-coordinates of the solution for (3.2), we compute the weekly incidence $\Delta\vec{S}_w = \vec{S}(7w-7) - \vec{S}(7w)$.

The likelihood of observing $f_{a,w}$ ILI cases in age class a during week w, is assumed to be Poisson distributed:

$$f_{a,w} \sim \mathrm{Poisson}(q_a \cdot [1-(1-\Delta S_{a,w})(1-B_{a,w})] \cdot g_{a,w})\,. \tag{3.3}$$

The Poisson distribution is taken because it approximates the Binomial distribution, but can be computed more efficiently, and has convenient additive properties (see below). The expectation is the product of the catchment population size $g_{a,w}$, the reporting probability q_a, and the probability of contracting either influenza ($\Delta S_{a,w}$), or another ILI-causing pathogen ($B_{a,w}$).

The incidence of background ILI ($B_{a,w}$) is modeled using the sine function, in order to capture seasonal effects [cf. 167]. We set $B_{a,w} = B_a(7w)$, where

$$\operatorname{logit}(B_a(t)) = b_{\text{age},a} + b_{\text{amp}} \sin\left(\tfrac{2\pi t}{365.24} - b_{\text{phase}}\right). \tag{3.4}$$

The age effect b_{age}, the phase $b_{\text{phase}} \in [0, 2\pi)$, and the amplitude $b_{\text{amp}} \in [0, \infty)$ of the background ILI have to be estimated.

The reporting probability q_a is allowed to follow a trend, by defining

$$\operatorname{logit}(q_a) = \mu_{\text{repo},a} + \alpha_{\text{repo},a} \cdot (s - \bar{s}). \tag{3.5}$$

The logit function on $(0, 1)$ is defined by $\operatorname{logit}(x) = \log(x/(1-x))$. The seasons ($s$) are centralized (by subtracting the middle season $\bar{s} = 22$ from s) to reduce the correlation between μ_{repo} and α_{repo}.

3.4.3 Estimation of the parameters

The parameters of the model are estimated using Markov chain Monte Carlo (MCMC). Since ODEs have to be integrated in order to compute the likelihood, we implemented a (Metropolis within) Gibbs sampler in `C++`. The code, together with mock data and a Python script for interpreting the output, has been made publicly available at `www.github.com/chvandorp/flu-sampler`. The costly likelihood computation can be accelerated by updating the seasonal sub-models in parallel, and using the fact that updating of many of the parameters does not require re-integration of the ODEs. The length of the chains is 2×10^5, and only the second half is used in the analysis, after applying a $1:100$ thinning. Convergence of the chain was assessed visually, by inspection of the trace plots.

The susceptibility parameters $S_{0,a}$ are given Logit-Normal prior distributions, dependent on season and age class a, with mean $\mu_{\text{susc},a} + \alpha_{\text{susc},a} \cdot (s - \bar{s})$ (where s denotes the season, and $\bar{s}$ the middle season) and variance $\sigma^2_{\text{susc},a}$, which are given non-informative hyper-priors $\mu_{\text{susc}} \sim \mathcal{N}(0, 100)$, $\alpha_{\text{susc}} \sim \mathcal{N}(0, 100)$ and $\sigma^2_{\text{susc}} \sim$ Half-Normal(100). In this way, susceptibility during seasons, or for age classes with little information, are informed by the other seasons and age classes. Similarly, the (relative) onsets of the epidemics t_0 are given a $\mathcal{N}(\mu_{\text{ons}}, \sigma^2_{\text{ons}})$ prior, with hyper-priors $\mathcal{N}(100, 1000) \backsim \mu_{\text{ons}}$ and Half-Normal(1000) $\backsim \sigma^2_{\text{ons}}$. The average background ILI $b_{\text{mean},a}$ is given a $\mathcal{N}(0, 100)$ prior, and the parameters governing the seasonality of the background ILI rate, b_{amp} and b_{phase}, are given a Half-Normal(100) and a Circular-Uniform$(0, 2\pi)$ prior, respectively. The offset μ_{repo} and slope α_{repo} of the logit-reporting rate both have a $\mathcal{N}(0, 100)$ prior distribution.

The proposal distribution for each parameter is given by a symmetric mixture of normal distributions [243], with mean equal to the present state of the parameter. The proposal variance is tuned during the burn-in phase to achieve an acceptance rate of about 0.44 [192]. For parameters whose domains have boundaries, the proposal is reflected in these boundaries to make sure that the Markov chain converges to the posterior distribution (due to the detailed balance condition). Likewise, a wrapped version of the proposal is used for circular domains.

3.4.4 Missing data

Some ILI data is missing (see the gray blocks in Figure 3.1), but part of this data could be retrieved from [241]. This recovered data is, however, not age-stratified, and also the catchment population sizes g are absent; instead the fraction of ILI incidence is reported. In order to compute the likelihood of observing a fraction of ILI incidence, given our model, we took the following steps. First, the time series of the present catchment population sizes shows a highly predictable pattern (Figure S3.2), and hence we took the simple approach of filling in the missing population sizes by taking a weighted average of the values in the 4 surrounding years:

$$(1-c_0)\hat{g}_{a,w} = \sum_{\substack{i=-2,\\ i\neq 0}}^{2} c_i g_{a,w+i\cdot 52} \tag{3.6}$$

with $c_{\pm 2} = 0.061$, $c_{\pm 1} = 0.245$, and $c_0 = 0.388$ (a Gaussian kernel). Second, suppose that μ_a is the expected ILI incidence (see Equation 3.3), and π is the reported fraction of ILI incidence, and f_a is the true ILI incidence, then $\sum_a f_a \approx \pi \sum_a \hat{g}_a \sim \text{Poisson}(\sum_a \mu_a)$, because of the additive properties of the Poisson distribution. The Gaussian kernel model does not correspond to the data that well between the years 2004 and 2006, but this poses no problem, since the data is age-stratified around that time.

3.4.5 Model selection

We compare the different models using the WAIC [74, 228], and the WBIC [229]. Models with a lower WAIC or WBIC are able to describe the data better. The WAIC, is an estimate of the out-of-sample prediction error. Ideally, one would use Bayes factors to compare different models, but computing Bayes factors is highly impractical in our case. The WBIC is acts as an approximation for the marginal likelihood (and hence, the difference in the WBIC for the Bayes factor). For non-singular models, one could replace WAIC by AIC, and WBIC by BIC. However, as our model is singular due to the use of hyper-parameters, we need to use WAIC and WBIC.

The WAIC equals $-2 \cdot (\widehat{\text{lpd}} - \hat{p}_{\text{WAIC}})$, where $\widehat{\text{lpd}}$ is the log point-wise predictive density, and $\hat{p}_{\text{WAIC}}$ is a penalty term based on the effective number of parameters [74]. Unlike the WAIC, the WBIC can not be computed from the output of the Gibbs-sampler. Instead, sampling has to occur at a different "temperature" determined by the number of observations ($n = 13\,333$). More precisely, $\text{WBIC} = -2 \cdot \mathbb{E}_\theta^\beta[L(D|\theta)]$, where $L(D|\theta)$ equals the log-likelihood of the data D given the parameters θ, and $\mathbb{E}_\theta^\beta$ denotes the expectation with respect to the distribution function proportional to $\theta \mapsto \exp(\beta L(D|\theta))\pi(\theta)$, with π denoting the prior distribution of θ, and $\beta = 1/\log(n) \approx 0.11$.

3.5 Supplementary material

3.5.1 Models with asymptomatic infection

$CD8^+$ T-cell memory might not result in reduced susceptibility, but does lead to less severe disease, and in the best case asymptomatic infection. Therefore, a model that includes an asymptomatic compartment would better fit with our assumptions about the epidemiology of influenza. However, as pointed out before for a simple case [167], such a model is equivalent to the usual SIR model, and the additional parameters are not identifiable.

Since we use a slightly more intricate model than the one explored by Van Noort *et al.* [167], we here show that also our model is equivalent with a version that includes asymptomatic compartments. This version of the model is given by the following ODEs:

$$\begin{aligned}
\tfrac{\mathrm{d}}{\mathrm{d}t}\vec{S} &= -\text{diag}(\vec{S})C(\vec{I}_1 + r\vec{A}_1 + \vec{I}_2 + r\vec{A}_2) \\
\tfrac{\mathrm{d}}{\mathrm{d}t}\vec{I}_1 &= (1-p)\cdot\text{diag}(\vec{S})C(\vec{I}_1 + r\vec{A}_1 + \vec{I}_2 + r\vec{A}_2) - 2\gamma\vec{I}_1 \\
\tfrac{\mathrm{d}}{\mathrm{d}t}\vec{I}_2 &= 2\gamma\vec{I}_1 - 2\gamma\vec{I}_2 \\
\tfrac{\mathrm{d}}{\mathrm{d}t}\vec{A}_1 &= p\cdot\text{diag}(\vec{S})C(\vec{I}_1 + r\vec{A}_1 + \vec{I}_2 + r\vec{A}_2) - 2\gamma\vec{A}_1 \\
\tfrac{\mathrm{d}}{\mathrm{d}t}\vec{A}_2 &= 2\gamma\vec{A}_1 - 2\gamma\vec{A}_2\,.
\end{aligned} \tag{3.7}$$

This model has two additional parameters: the reduction of infectiousness of the asymptomatic individuals in the A compartment (r), and the fraction of susceptible individuals that move to the A compartment upon infection (p). We have added 2 compartments A_1 and A_2 (of size A_1 and A_2, resp.) such that the asymptomatic phase has the same duration as the infectious phase.

When $r = 0$, individuals in compartment A are not infectious (possibly due to largely reduced viral shedding). In this special case, we can simply ignore the equations for $\vec{A}_i$. We will not be able to estimate the parameter p. Instead, we can write $\vec{S}^* = (1-p)\vec{S}$,

and can write the system of ODEs in terms of $\vec{S}^*$, and $\vec{I}_i$. The variable $\vec{S}^*$ can then be interpreted as the fraction of individuals that are susceptible for *symptomatic* infection. For general $0 \leq r \leq 1$, we can write $\vec{I}_i^* = \vec{I}_i + r\vec{A}_i$, and define $\vec{S}^* = (1-p+rp)\vec{S}$, so that

$$\begin{aligned}
\tfrac{\mathrm{d}}{\mathrm{d}t}\vec{S}^* &= -\mathrm{diag}(\vec{S}^*)C(\vec{I}_1^* + \vec{I}_2^*) \\
\tfrac{\mathrm{d}}{\mathrm{d}t}\vec{I}_1^* &= \mathrm{diag}(\vec{S}^*)C(\vec{I}_1^* + \vec{I}_2^*) - 2\gamma\vec{I}_1^* \\
\tfrac{\mathrm{d}}{\mathrm{d}t}\vec{I}_2^* &= 2\gamma\vec{I}_1^* - 2\gamma\vec{I}_2^*,
\end{aligned} \tag{3.8}$$

and these equations are identical to system (3.2). It is not unreasonable to assume that individuals in the A compartment have a reduced probability of consulting their GP compared to individuals in the I compartment. Let us denote the reporting probabilities q_A and q_I, respectively. The rate of reported ILI is therefore $(q_\mathrm{I}(1-p) + q_\mathrm{A}p)\frac{d}{dt}\vec{S}$. If we further assume that the probability of reporting is proportional to the reduction in infectiousness, we get that $(q_\mathrm{I}(1-p) + q_\mathrm{A}p)\frac{d}{dt}\vec{S} = q((1-p)+rp)\frac{d}{dt}\vec{S} = q\frac{d}{dt}S^*$. Hence, under these assumptions, the baseline model (3.2) is equivalent to a model with asymptomatic infection, but with S interpreted as susceptibility weighted by the likelihood of seeking health-care.

3.5.2 Models with frailty

In the case of SIR-type models, the term "frailty" refers to heterogeneity in susceptibility. The two extremes are zero-one susceptibility (a person is either fully susceptible to infection or not at all) and uniform susceptibility (everyone has the same (reduced) susceptibility).

An example of a model that incorporates both of these susceptibility notions is given by the following equations:

$$\begin{aligned}
\tfrac{\mathrm{d}}{\mathrm{d}t}\vec{S} &= -\mathrm{diag}(\vec{S})\mathrm{diag}(\vec{p})C\vec{I} \\
\tfrac{\mathrm{d}}{\mathrm{d}t}\vec{I} &= \mathrm{diag}(\vec{S})\mathrm{diag}(\vec{p})C\vec{I} - \gamma\vec{I},
\end{aligned} \tag{3.9}$$

with initial condition $\vec{S}(-\infty) = \vec{S}_0$. Here, the parameter $\vec{p}$ denotes the reduced susceptibility of the different age classes. Let $\vec{S}^* = \mathrm{diag}(\vec{S}_0)^{-1}\vec{S}$, and $\vec{I}^* = \mathrm{diag}(\vec{S}_0)^{-1}\vec{I}$, then

$$\begin{aligned}
\tfrac{\mathrm{d}}{\mathrm{d}t}\vec{S}^* &= -\mathrm{diag}(\vec{S}^*)\mathrm{diag}(\vec{p})C\,\mathrm{diag}(\vec{S}_0)\vec{I}^* \\
\tfrac{\mathrm{d}}{\mathrm{d}t}\vec{I}^* &= \mathrm{diag}(\vec{S}^*)\mathrm{diag}(\vec{p})C\,\mathrm{diag}(\vec{S}_0)\vec{I}^* - \gamma\vec{I}^*
\end{aligned} \tag{3.10}$$

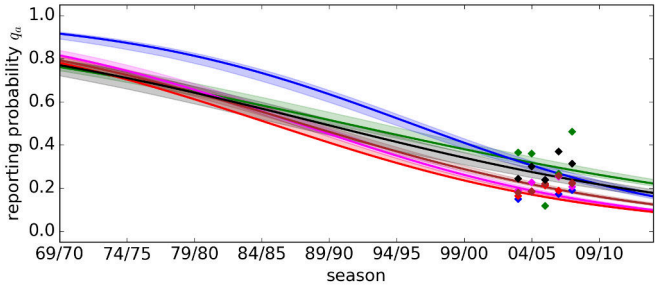


Figure S3.1: The reporting probability decreases with time. The reporting probabilities are given as a function of time (lines for the MAP estimate, and bands for the 95% CrI). The color coding for age class is identical to Figure 3.3. Notice that during a season the reporting probability is kept constant. The diamonds give the estimates based on the GIS data (cf. Figure 3.2).

The force of infection acting on age class a, is given by

$$\lambda_a = p_a \left(C_{a1} S_{0,1} I_1^* + \cdots + C_{a6} S_{0,6} I_6^* \right) . \tag{3.11}$$

Above, we interpret the parameters $S_{0,a}$ as the fraction of susceptible individuals (by using $\vec{S}_0$ for the initial condition). Equation (3.11) shows that $S_{0,a}$ can equally well be interpreted as reduction in infectiousness. In the case of immunity due to neutralizing antibodies, the best interpretation of $S_{0,a}$ would be the fraction of susceptible individuals. For cellular immunity, $S_{0,a}$ can probably be best interpreted as reduced infectiousness of individuals in age class a. Hence, in general, $S_{0,a}$ should be interpreted as a combination of both.

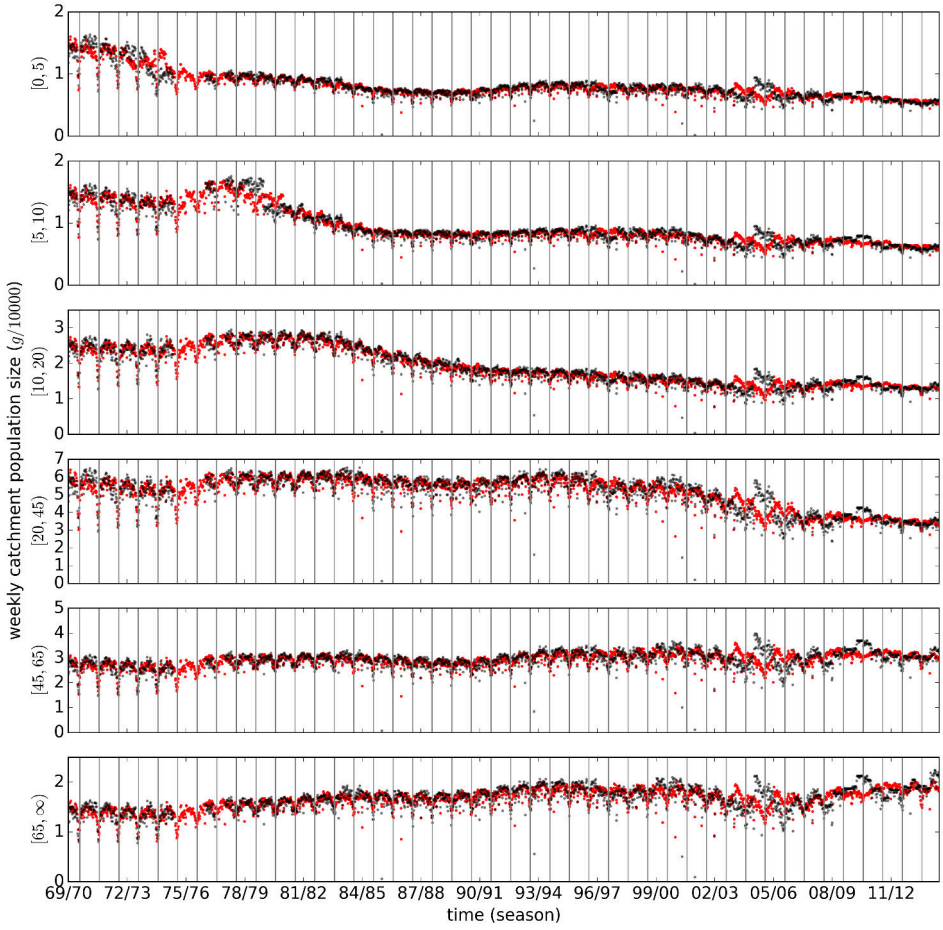


Figure S3.2: Catchment population size. The actual population sizes g are shown in black. The estimates ($\hat{g}$; Equation 3.6) based on the 4 surrounding years are shown in red. Only when the population size is not known, an estimate is used instead. The vertical lines show the beginning of the seasons (week 30).

Figure S3.3: Distribution of isolate sampling dates. The distribution of known isolate sampling calendar days is shown as a black histogram. The red line represents the smoothened density function.

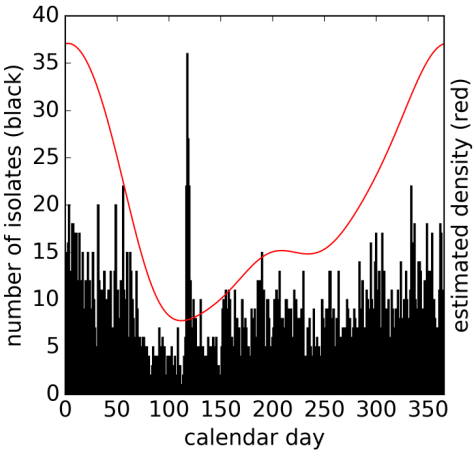
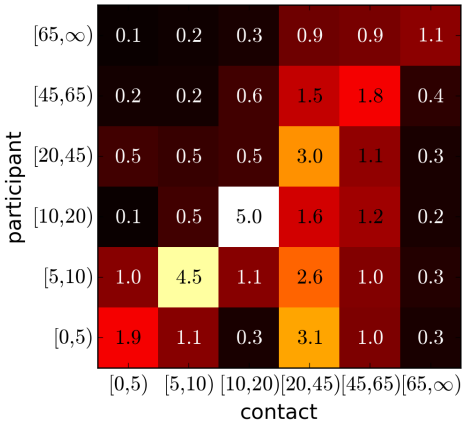


Figure S3.4: Contact matrix (C). The values represent the average number of contacts per day. The dominant eigenvalue of C equals 7.1.



Part II

A chronic infection: HIV-1

Chapter 4

Immuno-epidemiological modeling of HIV-1 predicts high heritability of the set-point virus load, while selection for CTL escape dominates virulence evolution

Christiaan H. van Dorp[1,2], Michiel van Boven[2], and Rob J. de Boer[1]

[1]Theoretical Biology and Bioinformatics, Universiteit Utrecht, Utrecht, The Netherlands.
[2]National Institute for Public Health and the Environment, Bilthoven, The Netherlands.

PLOS Computational Biology (2014)

Abstract. It has been suggested that HIV-1 has evolved its set-point virus load to be optimized for transmission. Previous epidemiological models and studies into the heritability of set-point virus load confirm that this mode of adaptation within the human population is feasible. However, during the many cycles of replication between infection of a host and transmission to the next host, HIV-1 is under selection for escape from immune responses, and not transmission. Here we investigate with computational and mathematical models how these two levels of selection, within-host and between-host, are intertwined. We find that when the rate of immune escape is comparable to what has been observed in patients, immune selection within hosts is dominant over selection for transmission. Surprisingly, we do find high values for set-point virus load heritability, and argue that high heritability estimates can be caused by the "footprints" left by differing hosts' immune systems on the virus.

4.1 Introduction

Human immunodeficiency virus type 1 (HIV-1) evolves under two levels of selection. On the one hand, there is within-host selection for immune escape. On the other hand, selection on the population-level acts on infectiousness and virulence. In this paper, we explore how these two levels of selection are intertwined, keeping in mind the massive heterogeneity of the hosts with respect to their cellular immune responses.

A HIV-1 infection can be separated into three phases: the acute phase, the asymptomatic phase and the symptomatic (or AIDS) phase. During the acute phase, the virus establishes high virus loads (the number of HIV-1 RNA copies per ml blood plasma) [129], until the $CD4^+$ target cells are depleted [175], and adaptive immune responses start limiting viral reproduction. The virus load then drops to a semi-stable level called the set-point. This marks the beginning of the asymptomatic or chronic phase, during which the partially restored $CD4^+$ T-cell count gradually drops, and at some point patients develop AIDS.

The set-point virus load (spVL) differs markedly between individuals. In untreated patients, spVL ranges from 10^2 to 10^6 copies/ml. The origin of this variation is an extensively researched topic, and explanations include host and viral factors. For instance, host factors incorporate the association between the set-point and the Human Leukocyte Antigen (HLA) haplotype, which is important for cellular immunity [57, 58, 123, 219]. The observation that the spVL is to some extent heritable [3, 86, 96, 98, 114, 160, 217, 245], suggests that viral genetic factors sway the set-point too. The exact extent of this heritability is unknown, as estimates range from 6% to 59%.

spVL is related to infectiousness and virulence. Patients with a higher spVL tend to be more infectious [230], but also develop AIDS more rapidly [149], resulting in a trade-off between infectiousness and the length of the asymptomatic phase. This life history trade-off was identified by Fraser *et al.* [65], and opens the door for HIV-1 adaptation with respect to transmission by means of spVL evolution. Certain spVL*s* (around $3.3 \cdot 10^4 \approx 10^{4.52}$ copies/ml) allow a HIV-1 strain to cause more secondary infections than strains with lower or higher set-points. A strain that establishes on average this optimal set-point should therefore become more abundant in the population. The striking observation is that, although large variation in set-points exists, most HIV-1 infected patients show a set-point close to the transmission-optimal value [65]. Moreover, mathematical models show that this adaptation can take place within realistic time scales [206], given the heritability estimates of spVL [98], and HIV-1's likely dates of origin [204, 239].

In such mathematical models, HIV-1's population-level fitness (measured in terms of the basic reproduction number $\mathscr{R}_0$) is only constrained by the life history trade-off, and environment- and mutation-induced spVL-variation. It is therefore quite intuitive that

in such a model evolution leads to intermediate levels of spVL [65, 206]. The inclusion of directed within-host evolution in such models introduces an extra constraint on the population-level fitness; one which dominates the evolutionary outcome, unless within-host selection is exceedingly weak. For a homogeneous host population, this has been shown recently by Lythgoe *et al.* [139], and they suggest that within-host evolution of traits affecting virus load must be slow. Below we argue that "short-sightedness" [*sensu* 126, 139], i.e., the life history trade-off has no apparent effect on the evolutionary outcome, can easily be understood when the host population is homogeneous. However, in a much more realistic situation where HIV-1 needs to escape from immune responses that vary markedly between individuals, the same intuition for the effect of directed within-host evolution can no longer be applied, and needs to be revised.

In this study, we explicitly incorporate such immune selection and massive host-heterogeneity with respect to immune responses in a nested epidemiological model. We investigate whether spVL evolution of HIV-1 is influenced by the virus' life history trade-off. Our model predicts that within-host immune selection has a major influence on population-wide spVL evolution. Thus, both Lythgoe's and our model predict short-sighted spVL evolution. However, we do not agree that within-host evolution must therefore be slow. Throughout the paper, we use the term "between-host adaptation" for evolutionary dynamics where HIV-1's life history trade-off notably affects the evolution of spVL. The term "within-host selection" refers to selection for immune escape and reversion of deleterious mutations.

At the same time, we use our model to investigate spVL heritability. We argue that high heritability can be a result of HIV-1 rapidly escaping immune responses, and the between-individual variation of these responses. We emphasize that spVL heritability caused by such a mechanism does not provide support for between-host adaptation.

4.2 Results

4.2.1 An immuno-epidemiological model

Our approach combines a caricature model for immune escape with a susceptible-infectious (SI) model for HIV-1 transmission. Both the within-host and the between-host simulations are discrete-event and individual based. The technical details are given in Methods. Here we present an intuitive exposition.

Cytotoxic T-Lymphocyte (CTL) responses are arguably important for controlling HIV-1 virus load [185, 201]. Human cells notify the cellular immune system about their proteome by presenting peptides on HLA molecules. On infected cells, a subset of these peptides originate from viral proteins. If a CTL clone detects such a foreign peptide,

symbol	description	value	note
k	size of a hosts' binding repertoire	$k \sim [\mathcal{N}(15,5)]$; $0 \leq k \leq n$	(1)
n	size of the union of all binding repertoires	300	(2)
e	number of escape mutations	$0 \leq e \leq k$	-
f	number of deleterious mutations	$0 \leq f \leq n-k$	-
m	the total number of mutations	$m = e + f$	-
$V_{\max}$	maximal $\log_{10}$ virus load	$4 \leq V_{\max} \leq 12$	-
$\lambda_{i,\text{esc}}$	escape rate in acute ($i = 1$), asymptomatic ($i = 2$) and AIDS ($i = 3$) phase	$\lambda_{2,\text{esc}} = 0.1 \cdot \lambda_{1,\text{esc}}$; $\lambda_{3,\text{esc}} = 0.5 \cdot \lambda_{1,\text{esc}}$; $10^{-2} \leq \lambda_{1,\text{esc}} \leq 10^{2}$	(3)
$\lambda_{i,\text{rev}}$	reversion rate during disease phase i	$\lambda_{i,\text{rev}} = r\lambda_{i,\text{esc}}$, where $r = 0.35$	(4)
V	$\log_{10}$ virus load	$V = V_0 - \sigma(k-e) - \varphi(e+f)$	(5)
σ	decrease in $\log_{10}$ virus load due to one immune response (without the fitness cost)	0.2	(6)
φ	fitness cost of a mutation	0.07	(6)
β_i	infection rate during disease phase i	$\beta_1 = 2.76\text{y}^{-1}$; $\beta_2(V) = \frac{\beta_{\max} 10^{V\beta_k}}{10^{V\beta_k} + \beta_{50}^{\beta_k}}$; $\beta_3 = 0.76\text{y}^{-1}$	(7)
D_i	(mean) duration of disease phase i	$D_1 = 0.24\text{y}$; $D_2(V) = \frac{D_{\max} D_{50}^{D_k}}{10^{VD_k} + D_{50}^{D_k}}$; $D_3 = 0.75\text{y}$	(7)

Table 4.1: Parameters and variables of the (standard) model. Notes: (1) $\langle k \rangle$ is chosen larger than observed numbers of immune responses [87, 130, 131], since we predict that viruses have escape mutations at infection, and do not escape all CTL responses. sd(k) is chosen to get reasonable variance in spVL, while limiting individuals with a very small binding repertoire. (2) about 10% of all possible peptides from HIV-1's proteome of ≈ 3000 a.a. (3) During the chronic phase, the escape rate slows down markedly [6, 87], hence we take $\lambda_{2,\text{esc}} = 10\% \cdot \lambda_{1,\text{esc}}$. The AIDS phase is sometimes preceded by escape from critical immune responses [171], and modeling suggests that escape rate speeds up towards the late disease phase [44]. Therefore we set $\lambda_{3,\text{esc}} = 50\% \cdot \lambda_{1,\text{esc}}$. (4) Both reports on fast [60] and very slow [69] reversion exist. We choose $r < 1$ in the order of magnitude of the ratio fitness cost and escape benefit. (5) The model for virus load was taken from Schmid *et al.* [200]. During the acute phase, V merely represents the virus fitness. *(continued on page 87)*

Table 4.1: *(continued from page 86)* Notes: (6) The magnitude $\sigma - \varphi$ is chosen to be in estimated ranges [102, 109]. Since escape appears to be faster than reversion, we choose $\varphi < \frac{1}{2}\sigma$. Although several studies find that a CTL response to Gag gives a $2-3$ fold higher fitness cost than $\varphi = 0.07$ [75, 109], we take φ as an average fitness cost. (7) The parameters β_i and D_i were taken from Fraser *et al.* [65]. The parameters for the Hill functions β_2 and D_2 are: $\beta_{\max} = 0.317\text{y}^{-1}$, $\beta_{50} = 13938\text{ml}^{-1}$, $\beta_k = 1.02$, $D_{\max} = 25.4\text{y}$, $D_{50} = 3058\text{ml}^{-1}$, $D_k = 0.41$.

it can kill the infected cell, and the peptide (in its proper HLA context) is called an epitope. Not all peptides can be presented by the HLA molecules of a host, and HIV-1 can escape from CTL recognition by mutating amino acids in its peptides to prevent presentation by the host's HLA molecules [80, 87, 130].

Due to HLA-polymorphism, the particular subset of all peptides that can be presented by a host's HLA molecules (the binding repertoire) differs strongly between individuals [82]. In our model we incorporate this by assuming that a wild-type virus has n peptides that can be presented in the population. A particular host can present a subset of size k of these n peptides. During infection, we assume that mutations in the n potentially recognized peptides occur according to a Markov process. Some of these mutations will result in CTL escape (escape mutations). In this case, the mutant takes over the viral population in that host. Naturally, if two hosts have have a common peptide in their binding repertoires, the mutated peptide is a CTL escape for both hosts.

In line with evidence, we assume that escape mutations in HIV-1 come with a fitness cost [75, 178]. The total fitness effect of an escape mutation, resulting from immune escape and its fitness cost, must be positive before the escape mutant can replace the dominant HIV-1 strain in the host. In order to model this, we use the virus load in the asymptomatic phase as a measure for within-host fitness. An immune response causes a reduction σ in the $\log_{10}$ virus load, and a fitness cost of any mutation reduces the $\log_{10}$ virus load by $\varphi < \sigma$. The total fitness effect of an escape mutation is then a $\sigma - \varphi$ increase in the $\log_{10}$ virus load. In the simulations, we choose $\sigma = 0.2$ and $\varphi = 0.07$ so that $\sigma - \varphi$ lies within estimated ranges [102, 109]. Qualitatively, our results do not depend on these particular choices for σ and φ, as long as $\varphi < \frac{1}{2}\sigma$ (results not shown).

Certain hosts have an efficient immune response to HIV-1. This can partially be explained by HLA-type. For instance, HLA-B*57, B*27, B*58 and B*18 are associated with a low spVL. HIV-1 is able to escape immune responses in hosts with these HLA-types, but the associated fitness costs tend to cripple the virus [130]. When such a crippled virus is transmitted to the next host, lacking the protective HLA-type, the virus load in this secondary host can remain low for a long time [75]. After a while, the crippled virus reverts the deleterious mutations, since the immune pressure causing these crippling mutations is not present in the secondary host [124]. We propose

that this effect is not only restricted to known protective HLA-types, but holds more generally [e.g., see 4]. We model this similar to immune escape. As a result of immune escape in previous hosts, a viral strain may carry a number of deleterious mutations. These mutations can revert to the wild-type, again according to a Markov process.

In summary, our model for the $\log_{10}$ virus load V is [cf. 200]

$$V = V_{\max} - \sigma(k-e) - \varphi(e+f)\,, \tag{4.1}$$

where $V_{\max}$ is the $\log_{10}$ virus load of a HIV-1 strain without deleterious mutations in the absence of CTL-responses ($k = 0$), e.g., the high virus load observed in a $CD8^+$ T cell depleted individual [73, 101, 201]. The integer e represents the number of escape mutations in a host (and hence, $k-e$ equals the number of immune responses), and f denotes the number of deleterious mutations. In other words, f equals the number of mutated peptides outside the current host's binding repertoire.

We assume that escapes and reversions appear at a rate proportional to the number of immune responses and deleterious mutations, respectively. Hence

$$e \xrightarrow{\lambda_{\text{esc}}\cdot(k-e)} e+1\,, \qquad f \xrightarrow{\lambda_{\text{rev}}\cdot f} f-1\,, \tag{4.2}$$

where λ_{esc} and λ_{rev} are the per-peptide rate of escape and reversion, respectively. We will refer to λ_{esc} and λ_{rev} as "mutation rates". Keep in mind, however, that our model of escape and reversion is quite phenomenological. The rates λ_{esc} and λ_{rev} are a combination of many factors, such as the error rate during reverse transcriptase and the fixation rate. Moreover, the rates λ_{esc} and λ_{rev} should in reality depend on the virus load. We simplify this dependence by assuming that the rates differ only between disease phases. In the acute and AIDS phase the per-peptide rates are high and in the asymptomatic phase, these rates are lower. Instead of λ_{esc} and λ_{rev}, we therefore take distinct parameters $\lambda_{i,\text{esc}}$ and $\lambda_{i,\text{rev}}$ for the per-peptide mutation rate in the acute ($i = 1$), asymptomatic ($i = 2$) and AIDS ($i = 3$) phase. We choose $\lambda_{i,\text{rev}} = r\lambda_{i,\text{esc}}$ with $r < 1$, meaning that reversion is slower than escape (see Table 4.1 for the exact parameterization). This is in line with the assumption that the total fitness benefit of an escape mutation is greater than the benefit of a reversion (i.e. $\varphi < \frac{1}{2}\sigma$).

As mentioned earlier, a HIV-1 strain infecting a new host carries a history of mutations acquired in previous hosts [75, 122]. In the context of the new host, many of these mutations will not be beneficial. Some of them may be advantageous, because HLA molecules can share epitopes [181, 182], and individuals share HLA molecules. To keep our model simple, we assume that a random host's binding repertoire is a random subset of size k of the set of all n possible HIV-1 epitopes. In reality, HLA haplotypes, and hence binding repertoires, are less regularly distributed. However, our simpler distribution provides us with the advantage that we only have to keep track of the *number* of

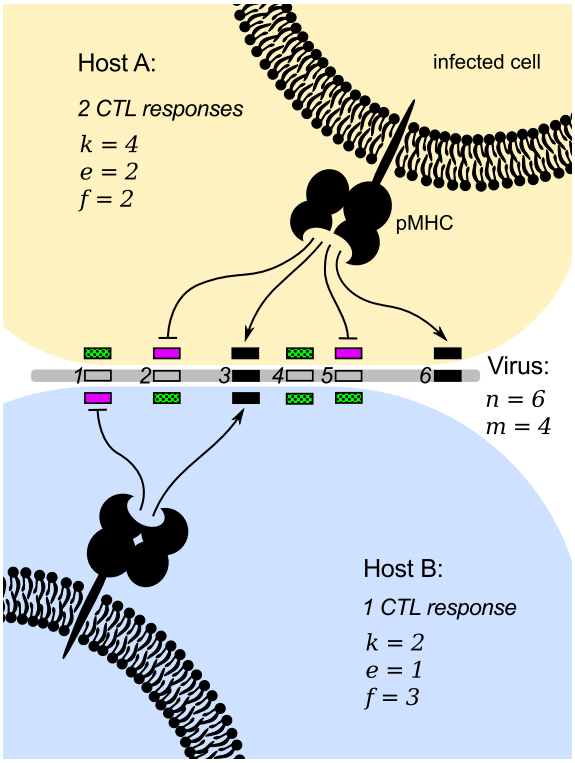


Figure 4.1: The phenotype of a virus differs between hosts, depending on the hosts' HLA haplotype. The virus in the figure has $n = 6$ potential epitopes (the rectangles), of which $m := e + f = 4$ have a mutation (the open rectangles). "pMHC" denotes the peptide-HLA complex. (Host A) Host A's HLA molecules can not bind peptides 1 and 4 (neither the wild-type, nor the mutant), but they can bind the wild-type of peptides 2 and 5. Thus, the purple rectangles denote immune escape mutations and the green (dotted) rectangles represent deleterious mutations. Since peptides 2 and 5 are mutated, they are escape epitopes in host A. The HLA molecules of host A can bind peptides 3 and 6, and hence peptides 3 and 6 are the epitopes for host A. During the infectious lifetime of host A, epitopes 3 and 6 may escape, and the mutated peptides 1 and 4 may revert to the wild-type. (Host B) The HLA molecules of host B bind less peptides of the wild-type virus ($k = 2$) than host A ($k = 4$); host B mounts a single CTL response against peptide 3. The HLA molecules of host B can also bind the wild-type of peptide 1, but this peptide is mutated, and hence peptide 1 is an escape epitope in host B. During host B's infection, epitope 3 may escape, and peptides 2, 4 and 5 may revert to the wild-type.

mutated peptides. Namely, when a host transmits a virus with e escape mutations and f deleterious mutations (denoted as an (e, f)-virus), then in the secondary host the virus will have phenotype (e', f') with $e + f = e' + f'$. We find the number of escape mutations e' by choosing a new random binding repertoire of size k'. Since every peptide is part of the new binding repertoire with equal probability, the number of *a priori* escape mutations is drawn from the hypergeometric distribution ($e' \sim \text{Hyper}(e + f, k', n)$). An example of how a virus' phenotype can differ between hosts is given in Figure 4.1. By default, we choose $n = 300$ and $k \approx 15$ (see Table 4.1), such that about 10% of HIV-1's peptides can serve as an epitope. The number k is chosen such that hosts have a realistic number of responses, also when many of the n peptides are mutated.

We model the three phases of a HIV-1 infection based on Fraser *et al.* [65] and Hollingsworth *et al.* [97]. The acute phase has a fixed length D_1, and in this phase individuals have a fixed infectiousness β_1. After D_1 years, the asymptomatic phase starts and infectiousness $\beta_2(V)$ and the average length of the asymptomatic phase $D_2(V)$ depend on the virus load V. The functions β_2 and D_2 are Hill functions with coefficients as estimated by Fraser *et al.* [65]. When the asymptomatic phase ends, the AIDS phase starts. This AIDS phase has, similar to the acute phase, a fixed length D_3 and fixed infectiousness β_3. We do not incorporate any correction for serial monogamy on infectiousness.

As an illustration of the within-host model, we have simulated a large number of within-host processes for two different parameter settings (Figure 4.2). Stochasticity and host-heterogeneity cause large variation in the within-host evolution of the virus (the thin step-wise lines in Figure 4.2). As deleterious mutations are reverted and CTL responses are escaped, the virus load increases during the infection. If the mutation rate is high, almost all escapes happen during the acute phase of the infection. The cohort-average virus load (the heavy blue line in Figure 4.2, bottom panels) can then even decrease, since individuals with a high set-point develop AIDS more rapidly. When these fast progressors die, we exclude them from the calculation of the cohort's mean virus load. Notice that during the acute phase, the variable V does not reflect the peak virus loads observed in patients, but is merely a measure of the virus' fitness.

For the between-host model, we explicitly model a population of infected individuals (of size I), and assume a frequency-dependent contact process with susceptible individuals [14]. Accordingly, super-infection and co-infection are ignored. We keep the total population size (N) constant, and only keep track of the susceptibles' number (S). Because of within-host evolution, an individual may transmit different viral strains during the course of an infection. When the virus load increases due to within-host adaptation, the infection rate also increases. We verified that a model with a non-constant population size does not give different results (not shown).

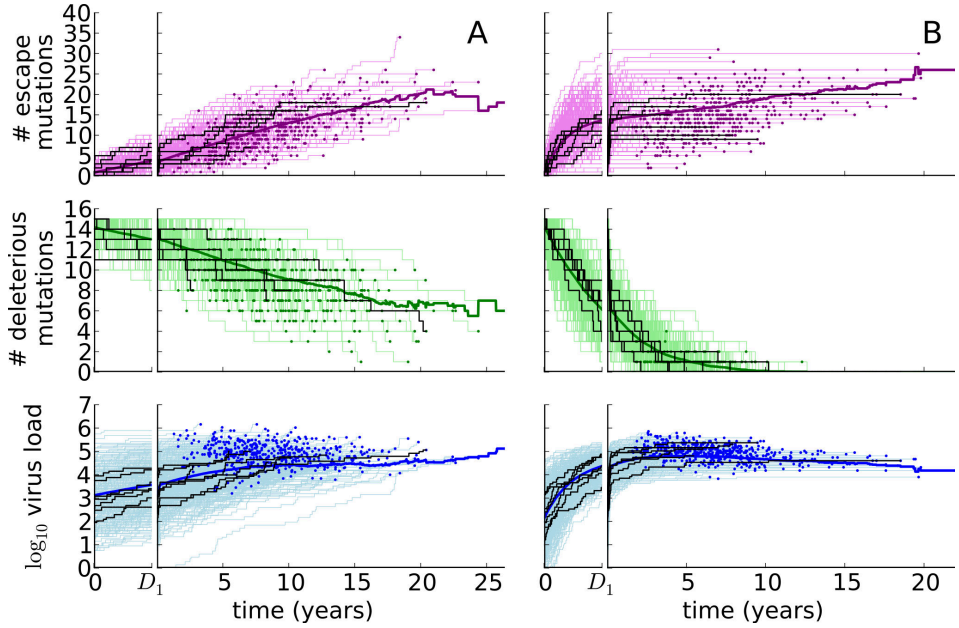


Figure 4.2: The within-host model for immune escape for different mutation rates. The graphs show the number of escape mutations (purple, top), the number of deleterious mutations (green, middle) and the virus load (blue, bottom). The mean number of mutations or virus load (the heavy lines) is based on 500 simulations (the thin step-wise lines). The dots indicate that a host died. All infections start with a virus with 15 mutations. The acute phase of the infection (the first $D_1 = 0.24$y) is displayed magnified on the left of each plot, and a couple of simulations are highlighted in black. **A.** The escape rate equals 10^0y^{-1}, and $V_{\max} = 7$. **B.** The escape rate equals 10^1y^{-1}, and $V_{\max} = 6$. Other parameters are listed in Table 4.1.

Since in our model the virus load can increase during the asymptomatic phase, we need to specify what we mean with set-point virus load. We define the spVL (in $\log_{10}$ scale) as the geometric average of the $\log_{10}$ virus loads in the asymptomatic phase, i.e., $\text{spVL} = \log_{10}\left(\frac{1}{L}\int 10^{V(t)}dt\right)$, where the integral is taken over the chronic phase, which lasts L years, and $V(t)$ denotes the virus load at time t. We often write $\langle\text{spVL}\rangle$ to indicate the population-wide arithmetic average spVL. Bracket notation is also used for other population-wide averages.

4.2.2 For realistic mutation rates, selection for immune escape dominates HIV-1 virulence evolution

When we choose the mutation rate low and run the agent-based model, the mean spVL converges to $4.52\log_{10}$ copies/ml; the value optimal for transmission (Figure 4.3A). However, this takes many centuries, depending on the maximal virus load V_{max} and the initial number of mutations. By increasing the mutation rate, we make the evolutionary dynamics faster, but lose between-host adaptation (Figure 4.3B). In fact, the mean spVL is approximately 1.3 $\log_{10}$ copies/ml higher than 4.52. By keeping the mutation rate equally high, but lowering V_{max}, the HIV-1 quasi-species can be given a population-level fitness ($\mathscr{R}_0$) that is about 17% higher than what is reached in Figure 4.3B. Apparently, selection for spVL values that are optimal for transmission is overruled by within-host selection at high mutation rates.

Both simulations in Figure 4.3 are approaching different steady states. Thus, to investigate between-host adaptation further, we now look at the properties of the model in population-level steady state for many different parameter combinations (Figure 4.4). To make the analysis computationally feasible, we stochastically approximate the next-generation matrix (NGM, see Methods). We fix all parameters except for the mutation rates ($\lambda_{i,\text{esc}}$ and $\lambda_{i,\text{rev}}$), and the maximum virus load (V_{max}). We keep the ratios $\lambda_{i,\text{esc}}/\lambda_{1,\text{esc}} = \lambda_{i,\text{rev}}/\lambda_{1,\text{rev}}$ and $r = \lambda_{i,\text{rev}}/\lambda_{i,\text{esc}}$ between the mutation rates constant (see Table 4.1 for the parameters chosen). Apart from the standard model described above, we also consider two modifications that serve as controls. In the first control, we take out the effect of population-level selection for transmission. In the second control, we make the population homogeneous.

The standard model. In this model, the population is heterogeneous ($n = 300$ and $\langle k\rangle = 15$), and virulence and infectiousness are taken from Fraser *et al.* [65], as described above. We refer to the resulting transmission potential ($\text{TP}(V) = \beta_2(V)\cdot D_2(V)$) as "peaked", because a single spVL value exists at which the number of secondary infections caused by one infected individual is maximal. Considering the population-level steady state for very high mutation rates can give us a measure of between-host

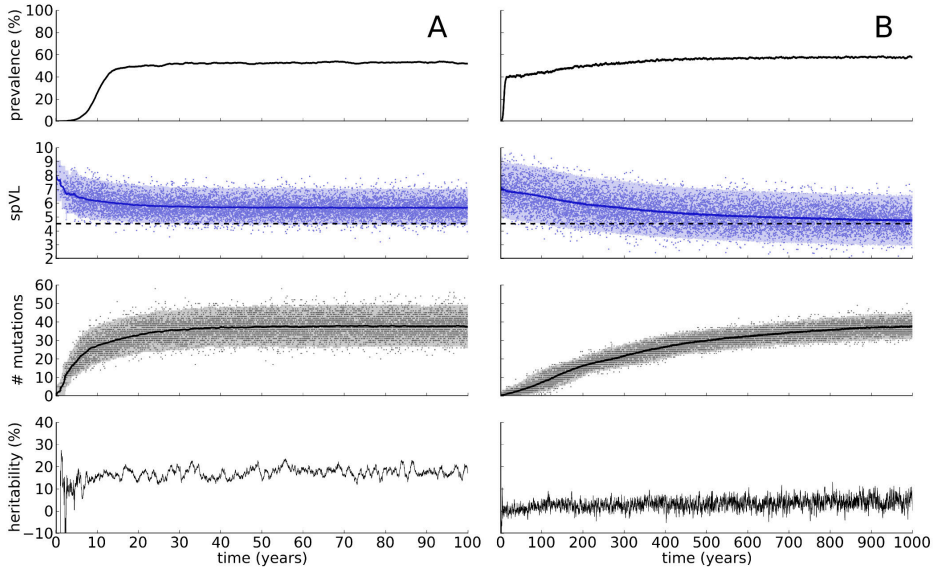


Figure 4.3: Two simulations of HIV-1 epidemics with two different mutation rates. The parameters are as follows: The maximal virus load equals $V_{\max} = 10$, and the population size equals $N = 25\,000$. **A.** The escape mutation rate in the acute phase equals $\lambda_{1,\mathrm{esc}} = 10^{-2}\mathrm{y}^{-1}$. **B.** The escape mutation rate in the acute phase equals $\lambda_{1,\mathrm{esc}} = 10^{0}\mathrm{y}^{-1}$. The other parameters are listed in Table 4.1. The simulations were started with 10 infected individuals that were infected with a virus with 0 mutations. The heavy lines in the graph of the set-point (spVL) and the number of mutations (# mutations) denote the population-wide average, i.e., $\langle \mathrm{spVL} \rangle$ and $\langle e + f \rangle$, respectively. The light bands denote the 2.5% − 97.5% percentiles, and the dots indicate the spVL of the receiver of a transmission couple (spVL) and the number of mutations of the transmitted strain (# mutations). In the graphs of the spVL, the dashed black line indicates the mean set-point that maximizes the transmission potential of HIV-1.

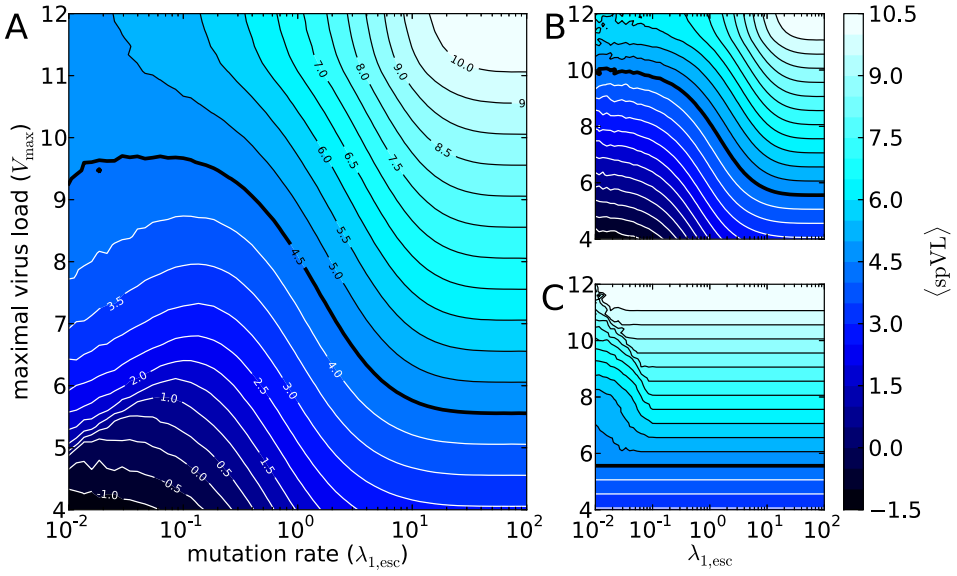


Figure 4.4: Exploration of the parameter space in three scenarios. The contours show the mean set-point virus load in the population-level equilibrium. The heavy black line indicates the graph of $V^*_{max}(\lambda_{1,esc})$, i.e., the value V_{max} for which $\langle$spVL$\rangle$ is optimal for transmission, given the mutation rate $\lambda_{1,esc}$. **A.** The standard model: A peaked TP and a heterogeneous host population. **B.** Control 1: A flat TP and a heterogeneous population. **C.** Control 2: A peaked TP and a homogeneous population.

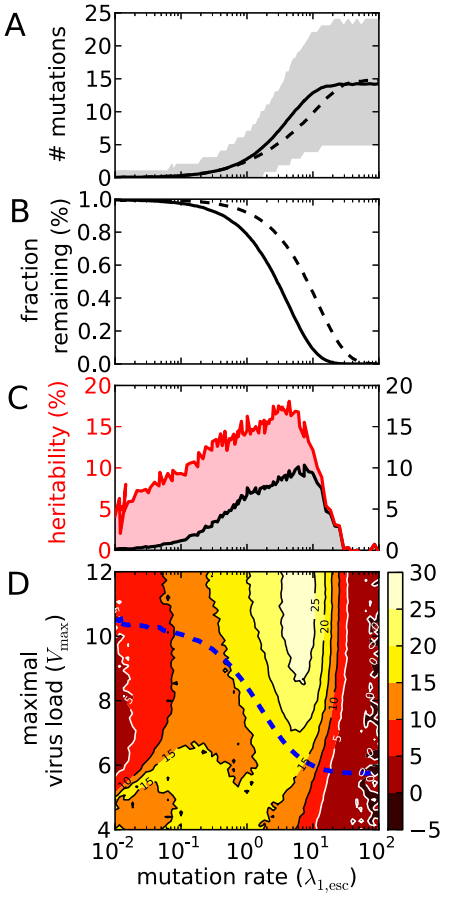


Figure 4.5: Heritability of set-point and the number of mutations during the acute phase. **A.** The average number of escape (solid) and deleterious (dashed) mutations at the end of the acute phase for different $\lambda_{1,esc}$. The resulting graph is barely dependent on V_{max} within the range $4 \leq V_{max} \leq 12$ (not shown). However, we choose V_{max} such that the mean set-point virus load ($\langle$spVL$\rangle$) equals 4.52, cf. the blue, dashed line in panel **D** (this also holds for panels **B** and **C**). The gray band indicates the $2.5\% - 97.5\%$ percentiles for the number of escape mutations in the acute phase. **B.** The mean fraction of immune responses (solid) and deleterious mutations (dashed) that remain after the acute phase. The resulting graph is barely dependent on V_{max} within the range $4 \leq V_{max} \leq 12$ (not shown). **C.** Heritability as a function of the mutation rate (upper red line). The black line below corresponds to the contribution of the immunological footprint to heritability, as estimated with the SEM. **D.** Heritability of set-point virus load for different combinations of V_{max} and $\lambda_{1,esc}$ for the standard model in steady state. The blue, dashed line indicates the contour where the $\langle$spVL$\rangle$ equals 4.52.

adaptation. When we choose the escape rate in the acute phase ($\lambda_{1,\text{esc}}$) close to 10^2y^{-1}, the virus will escape all immune responses, and revert all deleterious mutations acquired in previous hosts, during the first few weeks of the infection (Figure 4.5B, black graphs). The population-average spVL will therefore tend to $\langle\text{spVL}\rangle \approx V_{\max} - \langle k\rangle\varphi$. In Figure 4.4A this relation is visible when $\lambda_{1,\text{esc}} > 10^1\text{y}^{-1}$ from the equidistant contours of the set-point in equilibrium. When we replace $V_{\max}$ by, say, $V_{\max}+0.5$ and we observe that $\langle\text{spVL}\rangle$ changes into $\langle\text{spVL}\rangle + 0.5$, then the virus is not capable of between-host adaptation. In our parameter space exploration, this isometric dependence of $\langle\text{spVL}\rangle$ on $V_{\max}$ can not only be observed for unrealistically high, but also for intermediate mutation rates.

Notice that our parameter space exploration can be regarded as a sensitivity analysis. For a fixed escape rate $\lambda_{1,\text{esc}}$, a value $V^*_{\max}(\lambda_{1,\text{esc}})$ exists such that $\langle\text{spVL}\rangle$ is optimal for transmission. By varying $V_{\max}$ around the value $V^*_{\max}$, we can study the sensitivity of $\langle\text{spVL}\rangle$ with respect to $V_{\max}$. In Figure 4.4 the graphs of $V^*_{\max}$ are given by heavy black lines.

A different picture emerges for low mutation rates ($\lambda_{1,\text{esc}} \approx 10^{-2}\text{y}^{-1}$). The contours are no longer equidistant, indicating that the steady state, the result of mutation and selection, is less sensitive to changes in $V_{\max}$. The absence of $V_{\max}$-sensitivity is most noticeable when $V_{\max} \approx V^*_{\max}$. This suggests that the virus is able to adapt on the level of the population (between-host adaptation). We will confirm this using Control 1 below. As mentioned before (Figure 4.3), between-host adaptation takes many centuries for small mutation rates. The within-host process for this parameter regime is extremely slow.

By considering the number of escape mutations in the acute phase, we can get insight into what parameter regime is realistic for HIV-1. Several studies show that the number of escape mutations in the first months after infection varies among patients, and lies between 1 and 10 [87, 130, 131]. This suggests that for escape rates $\lambda_{1,\text{esc}}$ to be considered realistic, they must be in the range 10^{-1} to 10^1y^{-1} (see Figure 4.5A). For these intermediate mutation rates, we see a strong effect of host-heterogeneity. Host-heterogeneity and subsequent infections that require new escape mutations, account for the accumulation of deleterious mutations, since deleterious mutations are not lost at a fast enough rate. The virus' inadequacy to fully adapt to individuals during infection decreases the within-host virus load. We will further justify this with Control 2 below. However, the lack of perfect adaptation to individuals' immune systems does not noticeably facilitate between-host adaptation. In the regime of realistic CTL escape rates ($10^{-1}\text{y}^{-1} \leq \lambda_{1,\text{esc}} \leq 10^1\text{y}^{-1}$), spVL evolution is not driven by the life history trade-off, as can be seen from the isometric relation between $V_{\max}$ and $\langle\text{spVL}\rangle$ (Figure 4.4A).

Control 1, eliminating population-level selection. By comparing the standard model with a model where no between-host adaptation is possible, we can study the impact of the peaked transmission potential. We eliminate selection for transmission by scaling each host's infectiousness such that the expected number of secondary infections during an entire infectious lifetime equals the same constant for all individuals. The transmission potential will therefore be "flat". To make this precise, let $\beta(t)$ denote an individuals infection rate at time t (depending on disease phase or virus load). We now want to make sure that each individual is expected to infect 2 new individuals (in a fully susceptible population). To achieve this, we replace $\beta(t)$ with $2 \cdot \beta(t) / \int_{t_{\text{infect}}}^{t_{\text{death}}} \beta(t')\mathrm{d}t'$. Here t_{infect} and t_{death} denote the time of infection and death, respectively. Notice that we do allow for variation in virulence; the length of the infectious period ($t_{\text{death}} - t_{\text{infect}}$) is equally dependent on virus load as in the standard model. Hence, we eliminate selection for transmission without altering the within-host process.

Figure 4.4B shows the mean spVL in steady state for the flat TP and a heterogeneous population ($n = 300$, $\langle k \rangle = 15$). The evolutionary outcomes for control 1 and the standard model are nearly identical when the escape rate $\lambda_{1,\text{esc}} > 10^{-1}\text{y}^{-1}$. This confirms that the virus is not capable of between-host adaptation when mutation rates allow for a realistic number of escape mutations during an infection. For extremely low mutation rates ($\lambda_{1,\text{esc}} < 10^{-1}\text{y}^{-1}$), we see a difference between the model with a flat and a peaked TP, confirming that between-host adaptation relies on low mutation rates.

In the slow mutation regime, optimizing the life-history trade-off can be accomplished in two ways. If the virus were to experience a flat transmission potential, then a quasi-species' spVL distribution and the number of accumulated mutations would only be determined by the rate of escape and reversion, and the heterogeneity of the host population. If the same species starts evolving under the influence of a peaked transmission potential, then the number of mutations might either decrease, resulting in a lower fitness cost, and a higher $\langle \text{spVL} \rangle$ (e.g., when $V_{\max} = 8$ and $\lambda_{1,\text{esc}} = 10^{-2}$), or the number of mutations might increase, resulting in a lower $\langle \text{spVL} \rangle$ (e.g., when $V_{\max} = 11$ and $\lambda_{1,\text{esc}} = 10^{-2}$). Notice that mutations always arise as CTL escapes in an individual, but that such a mutation is most often deleterious in the other hosts.

Control 2, the homogeneous case. The effect of HLA-polymorphism can be studied by considering a model without host-heterogeneity. In the models with heterogeneous host populations, we assigned upon infection of a new individual a random binding repertoire, i.e., a random subset of size k of the virus' n potential epitopes. For this control, we assign to every host exactly the same binding repertoire. As a consequence, escape mutations remain beneficial after infection of a new host. Notice that in our model deleterious mutations always originate from escape mutations in earlier hosts

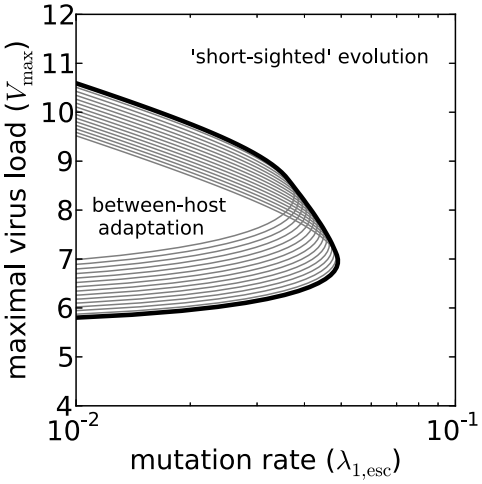


Figure 4.6: The bifurcation in the homogeneous model. For each $\ell = 0, \ldots, k-1$ a thin gray line indicates the curve $\{G_{\ell,\ell} = G_{k,k}\}$. The heavy black line separates the region of the parameter space (between-host adaptation) where $G_{k,k} \geq G_{\ell,\ell}$ for all ℓ.

with a non-identical binding repertoire. Therefore, deleterious mutations are purged from the population.

Figure 4.4C shows the mean spVL in steady state in case of a homogeneous host population (with a peaked TP). For a wide range of mutation rates, the virus manages to escape all immune responses. As a consequence, there is no room for population-level adaptation with respect to transmission, which is visible from the equidistant $\langle \text{spVL} \rangle$-contours, and the isometric relation between $V_{\max}$ and $\langle \text{spVL} \rangle$. If we compare Figure 4.4A and C at intermediate mutation rates ($10^{-1}\text{y}^{-1} \leq \lambda_{1,\text{esc}} \leq 10^{1}\text{y}^{-1}$), we indeed see that host-heterogeneity lowers the set-point drastically.

For small mutation rates ($\lambda_{1,\text{esc}} \leq 10^{-1}\text{y}^{-1}$) we observe a threshold, which depends on $V_{\max}$ and $\lambda_{1,\text{esc}}$. For mutation rates below this threshold, viruses evolve that do not escape all immune responses. When the mutation rate is small enough, escape mutations are rare. Viral strains that do escape yet another immune response will establish a higher set-point virus load. These escape mutants are then out-competed on the population level by strains that are better recognized by CTLs, because the life-history trade-off favors a lower set-point.

The above mentioned threshold can be well understood by studying the NGM. The stochastic model can be simplified so that a mathematical analysis is possible. The threshold for the homogeneous model can be described in terms of the eigenvalues of the NGM and is caused by a transcritical bifurcation (see Methods). The heavy black line in Figure 4.6, which we find with the mathematical analysis, gives the location of the bifurcation. This line separates the parameter space into a region where between-host adaptation is possible, and where it is not. The line coincides with the apparent threshold that can be observed in Figure 4.4B.

4.2.3 Immune escape causes heritability of set-point virus load

Between-host adaptation is only possible if the spVL is inherited from one person to the next. If the speed of within-host adaptation is intermediate or fast, our model does not predict population-level adaptation for transmission. In order to verify that the absence of between-host adaptation is not due to lack of spVL heritability (h^2, see Methods), we compute heritability during an epidemic (see Figure 4.3, bottom panels), and in the steady state of the (standard) model for many different parameter combinations (see Figure 4.5C).

During a simulated epidemic, we use all transmissions that take place within a time span of a year to compute heritability. This means that the sample size for computing heritability equals the (yearly) incidence. The median incidence in the simulation with a low mutation rate ($\lambda_{1,\text{esc}} = 10^{-2}\text{y}^{-1}$, Figure 4.3B) equals 2335 yearly infections (95% CI: $[1839, 4085]$). For the simulation with a faster mutation rate ($\lambda_{1,\text{esc}} = 1\text{y}^{-1}$, Figure 4.3A), more virulent viruses evolve, and the median incidence equals 3209 infections per year (95% CI: $[241, 3586]$). Even with these large sample sizes, heritability fluctuates substantially over time. In Figure 4.3B the median h^2 equals 3.12% (95% CI: $[-1.31, 8.29]$), and in Figure 4.3A the median h^2 equals 17.1% (95% CI: $[10.8, 21.8]$). The rapid fluctuation in h^2 might explain why different experimental studies to HIV-1 spVL heritability that use transmission couples [86, 98, 114, 217, 245] often give quite varying results [cf. 160]. The NGM approach allowed us to produce an even larger number of transmission couples, and hence, to estimate heritability more accurately. Overall, heritability lies between 0% and 30%, and is $\geq 10\%$ for realistic parameter combinations.

In our model, one can think of two mechanisms that cause heritability. The first mechanism applies when mutation rates are not too high. If variation in the number of mutations exists and the mutation rate is low, the spVL of transmitting and recipient hosts are correlated, although this correlation will not be perfect due to the variation in the breadth of the immune response (k). If the mutation rate increases, viruses adapt to their host more rapidly and, according to this first mechanism, the correlation vanishes.

The second mechanism is related to transmission of crippled viruses. If a host controls the infection well because of a broad immune response, the virus will escape more CTL responses and, when transmitted, becomes crippled in the average new host. In the primary, controlling host, the set-point virus load is low due to good initial immune responses and the virus' fitness cost of escape, and in the average secondary host the virus load will again be low due to the high number of deleterious mutations. *Vice versa*, in hosts with a narrow immune response, transmitted strains will have few new escape mutations and this will lead to few deleterious mutations in the recipient.

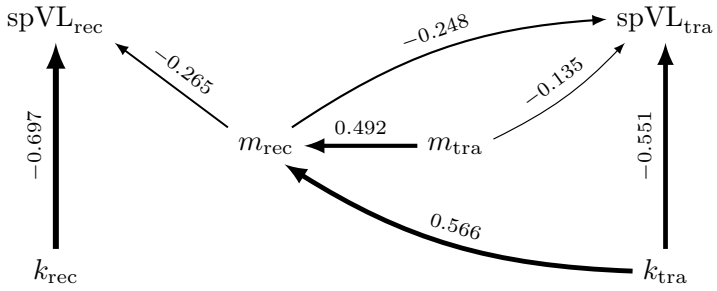


Figure 4.7: The structural equation model (SEM) used for quantifying the immunological footprint. Shown is a directed, acyclic graph (DAG) representing the SEM. The arrows indicate dependencies between the variables. The numbers above the arrows are the fitted weights (all highly significant: $p < 10^{-3}$), and the size of these weights is also represented by the thickness of the arrows. The data for this example comes from a simulation of the standard model, with $\lambda_{1,\text{esc}} = 3\text{y}^{-1}$ and $V_{\text{max}} = 6.64 \log_{10} \text{ml}^{-1}$.

We can most clearly see the effect of the second mechanism when both mutation rate and V_{max} are high (the contour $h^2 = 25\%$ in Figure 4.5C). In this part of the parameter space, most immune responses are escaped in the acute phase (cf. the solid graph in Figure 4.5B). Rapid escape causes variation in the number of deleterious mutations in the transmitted virus, because the size of the binding repertoire (k) varies among individuals. However, when $\lambda_{1,\text{esc}} = \lambda_{1,\text{rev}}/r < 30$, not all deleterious mutations can be reverted in the acute phase (cf. the dashed graph in Figure 4.5B). For high V_{max}, the asymptomatic phase is short, resulting in few reversions during this phase and a 'footprint' of the transmitting host's immune responses on the receiving host's spVL [245]. Notice that the second mechanism does depend on the reasonable assumption that reversion is a slower process than escape ($\lambda_{i,\text{esc}} > \lambda_{i,\text{rev}}$ and $\varphi < \frac{1}{2}\sigma$, not shown), and that the size of the binding repertoire (k) differs between individuals.

As the above evidence for the second mechanism—or "footprint effect" as we like to call it—is only circumstantial, a quantification of this mechanism is needed. To quantify the footprint effect we analyze the simulations using a structural equation model (SEM). The model estimates heritability, and takes the fitness costs ($m = e + f$) and breadth of the immune response (k) into account. Heritability of spVL is the sum of two effects; one mediated by viral fitness, and the other by the breadth of the immune response of the transmitting host. Figure 4.7 shows a graphical representation of the model, and details of the analyses are given in the Methods section.

For realistic parameter values, approximately half of the observed heritability is due to the footprint effect (Figure 4.5C). When we lower the rate of escape, the footprint

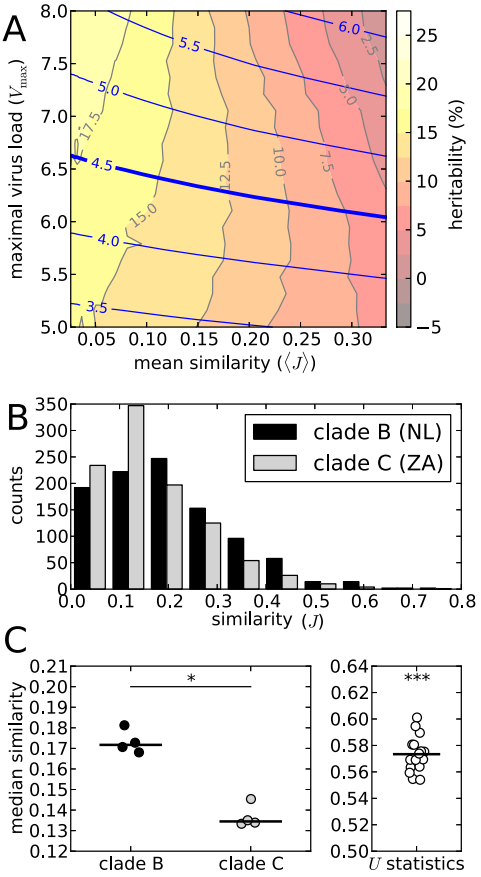


Figure 4.8: Host-heterogeneity and heritability. A. The panel shows a contour plot of heritability (h^2, gray lines, red/yellow faces) as a function of the maximal virus load (V_{max}) and the expected similarity between binding repertoires ($\langle J \rangle$). On top of the heritability contour plot, the blue lines indicate the contours of $\langle$spVL$\rangle$. The heavy blue contour corresponds to the transmission-optimal $\langle$spVL$\rangle$. **B.** Distributions of the overlap between pairs of binding repertoires. The black bars correspond to European HLA-haplotypes and a clade B virus (sampled in the Netherlands). The gray bars correspond to Sub-Saharan HLA-haplotypes and a clade C virus (sampled in South Africa). The distributions were simulated by sampling a 1000 HLA-haplotype pairs. **C.** Statistics on the sampled distributions as in panel **B**. The left panel shows the medians of the similarity distributions for 4 strains representative of clade B (black dots) and clade C (gray dots). The difference is significant (Mann-Withney U-test, $p < 0.05$, *). The right panel depicts the U-statistic for all clade B and clade C pairs. The mean of the U-statistics is significantly larger than 0.5 (t-test $p < 0.001$, ***).

effect, and therefore also the heritability, decreases. On the contrary, when within-host evolution is extremely fast, almost all of the heritability is due to the footprint effect, although the total heritability decreases.

4.2.4 Host heterogeneity and spVL heritability: A model-based prediction

Our model predicts that heritability of the set-point virus load and host-heterogeneity are related. When within-host evolution is fast enough, approximately half of the observed heritability may be explained by the immunological footprint. Also, when we lower heterogeneity in our model, heritability will decrease.

An intuitive measure for heterogeneity in the host population is the expected similarity of hosts' binding repertoires. This tells us how much adaptation to one host remains beneficial in the next. As a measure of the similarity of two binding repertoires K_1

and K_2 (of size k_1 and k_2, respectively) we use the Jaccard index $J := |K_1 \cap K_2|/|K_1 \cup K_2|$, the overlap between binding repertoires, divided by the the number of (wild-type) epitopes that at least one of the hosts can recognize.

Figure 4.8A shows the relation between the expected similarity between hosts ($\langle J \rangle$) and the heritability of the set-point (h^2). We modulated heterogeneity by varying n, the total number of potential epitopes, between 30 and 300, corresponding to low and high host-heterogeneity, respectively. The mutation rate $\lambda_{1,\text{esc}}$ equals 3y^{-1}, such that the number of escape mutations during the acute phase lies within a realistic range. The figure shows that heritability indeed decreases when the population becomes more homogeneous, which indicates that high heritability relies on host-heterogeneity.

HLA-heterogeneity differs between human populations. If our model prediction holds, then this variation could affect the heritability of the set-point measured in these populations. An unpublished study by Hodcroft *et al.* [95], shows that heritability in measured for HIV-1 clade C in a Sub-Saharan African population is higher than heritability for HIV-1 clade B in a European cohort [96] (30% vs. 5.7%). Keeping our model in mind, we are able to understand this, if the European population with respect to clade B, is less heterogeneous than the Sub-Saharan population with respect to clade C.

Using the peptide-MHC binding predictor NetMHCpan [99], we compared the two populations and circulating viruses (see Methods). Again, we measured heterogeneity in terms of similarity between binding repertoires. We sampled from the HLA-haplotype distributions of the European and Sub-Saharan populations, and calculated how similarity (again measured in terms of the Jaccard index) within the populations is distributed. Figure 4.8B shows two of these distributions. The black bars correspond to the European population, and the gray bars to the Sub-Saharan population. Although small, these populations do show a difference in heterogeneity: The Sub-Saharan population is more heterogeneous than the European population, as European binding repertoires tend to be more similar. The difference in heterogeneity is statistically significant (see Methods and Figure 4.8C).

4.3 Discussion

In this paper, we model HIV-1 transmission and within-host adaptation by means of immune escape in a HLA-heterogeneous host population. In comparison to what data suggests, we do not find that HIV-1's life history trade-off determines or influences spVL evolution. For realistic mutation rates, the evolutionary outcome is mostly determined by within-host selection for escape and reversion. Without HLA-heterogeneity, viruses would evolve to be within-host optimal in every host. Due to HLA-polymorphism,

however, deleterious mutations accumulate, and the environment changes at each transmission. This causes virulence to evolve to intermediate levels for most hosts. Incomplete adaptation at the individual level is not exploited by the virus in order to improve its transmission potential. Although set-point virus loads are expected to be lower in a heterogeneous population, spVL evolution remains short-sighted. As we will point out below, our model is limited in the sense that we only incorporate immune escape and reversion as a means for within-host and between-host adaptation. Nevertheless, since population-level adaptation does occur when within-host adaptation is slow, the model's limitations do not necessarily revoke our conclusions.

In our model, we do find that spVL is heritable, even when the mutation rate is high. spVL heritability is needed for between-host adaptation. However, for realistic mutation rates, high heritability, as measured using transmission couples, is over-estimated; it mostly results from a footprint left by the transmitter's immune system on the receiver's spVL. This novel explanation calls the validity of the use of high heritability as support for between-host adaptation in question. During real HIV-1 infections, immune escape sometimes requires compensatory mutations. Such escape variants need more time to revert to the wild-type in hosts lacking the escaped CTL response [44]. Such a mechanism is not incorporated in our model, but is likely to cause even higher heritability compared to what we find. Given previous results on the effects of transmitted CTL escape mutations on a receiver's virus load [75, 124], and the sharing of HLA alleles [53, 245], we think the footprint effect provides a sound explanation for the experimentally observed high heritability of the set-point. Importantly, if this explanation were to be found true, and if spVL evolution and heritability are indeed strongly influenced by CTL escape, reversion and compensatory mutations, finding SNPs in HIV-1's genome that control spVL might be a fool's errand, unless this pursuit would be restricted to known CTL-epitope sites.

Our claims concerning the footprint effect, and the dependency of heritability on host-heterogeneity are not just speculative. We show that the model can make testable predictions, and we give an example of how such a test can be performed, i.e., by comparing host-heterogeneity in different human populations. In our example, we compared Sub-Saharan and European populations with respect to the viruses circulating in these populations, and showed that host-heterogeneity is higher in the African population, which is consistent with our novel explanation, and estimates of the heritability in these populations. Of course, we would not suggest that this isolated finding is evidence for the footprint effect, although we do want to stress that heritability estimates are expected to be correlated with host heterogeneity. Moreover, the heritability estimates that were used in this example were obtained using a phylogenetic analysis [95, 96], while our explanation only holds for studies that use transmission couples. In future work, we plan to investigate whether an immunological footprint can also affect heritability that has been estimated using phylogenies or pedigrees.

Intuitively, the fact that within-host adaptation overrules between-host adaptation can be understood by considering that many viral generations separate the founding virus and a transmitted strain, while transmission only takes one generation. In the homogeneous model, this results in full within-host adaptation (throughout the population all epitopes are escaped), except when within-host adaptation is extremely slow. This result was also shown recently by Lythgoe *et al.* [139].

The intuition mentioned above works best for homogeneous populations. Adaptations to a primary host are beneficial again in a secondary host, and if within-host adaptation is fast, this leads to population-wide within-host adaptation and not between-host adaptation. This part of the intuition fails for a heterogeneous host population, where within-host adaptations in the form of immune escapes, are most likely not beneficial in the next host. Therefore, one could argue that homogeneity obstructs between-host adaptation. Here, we attempt to remove that obstruction by adding host-heterogeneity to a multi-level HIV-1 model. We find that in a heterogeneous population, HIV-1 also fails to evolve a mean spVL that maximizes the transmission potential, as shown by our sensitivity analysis and controls. Of course, when we make within-host adaptation trivial by choosing a very low mutation rate, population-level adaptation occurs.

Apparently, host-heterogeneity does not solve the within- versus between-host adaptation paradox. Our models tell us that within-host adaptation overrules between-host adaptation, and yet HIV-1 appears to have adapted with respect to the life history trade-off [65], or at least is evolving its mean spVL towards the value that maximizes the transmission potential [90]. Several mechanisms that can serve as a solution for the paradoxical observation have been proposed [66, 100, 137].

One of these mechanisms is referred to as "store and retrieve" [137]. It is hypothesized that latently infected memory $CD4^+$ cells occasionally produce virus, and that these virions are preferentially transmitted. Preferential transmission is backed up by the observation that evolutionary rates are higher at the within-host than at the between-host level [1], and recently by a very interesting study into HIV-1's transmission bottleneck [30]. However, transmission of CTL escape mutants within transmission couples [75], and even the spread of CTL escape mutants through populations has been observed [51, 106, 122, 154, 199]. These observations indicates that 'store and retrieve' is not absolute, and in order for this mechanism to solve the paradox, we expect it to rely on getting the population-level evolutionary rate below a threshold; one which may not be reached. This premise could limit the robustness of the "store and retrieve" model. Furthermore, when the population-level evolutionary rate is slowed down because of a mechanism like "store and retrieve", the rate of between-host adaptation is also decreased, which could conflict with the short time scales at which adaptation must have been taking place for HIV-1 [204, 206].

Another possible mechanism is a heritable viral trait that influences spVL, but that is not under within-host selection. This was recently examined by Hool et al. [100]. An example of such a trait could be target cell activation rate [9, 198]. In short, if a viral trait influences target cell activation, and a mutant strain manages to increase the activation rate, then this additional activation is a common good for the entire within-host viral population (activated cells produce more virions). Hence, the mutant does not have an advantage and will not be preferentially selected. Drift creates within-host variation in activation rate, and the transmission bottleneck leads to variation of the target cell activation rate at the population-level. This hypothesis could be challenged by other traits that affect spVL, since these may still be under within-host selection, and are likely to interfere with the within-host neutral one.

We finish with a novel suggestion for solving the paradox, one which is based on our modeling formalism, and was recently also put forward by Fraser *et al.* [66]. One point of criticism on our model could be that we limit the evolutionary capabilities of our *in-silico* viruses. Strains can only evolve their number of deleterious mutations in order to approach population-level favorable spVLs. Unfortunately, in the current framework, it is not sensible to allow for mutations in other parameters, in particular V_{max}, since then V_{max} would only increase during within-host evolution, and hence, during the course of the epidemic. This is because we assume that no two strains simultaneously reside a single host, and that mutants with a higher fitness go to fixation rapidly. In reality, fixation of mutants within a host can take a considerable amount of time [6].

An obvious—but technically challenging—fix for this problem is to abandon the assumption that the within-host evolutionary dynamics is memoryless, and allow for multiple mutants to compete for fixation, i.e., allow for clonal interference [107, 118, 127, 172, 216]. These mutants can then carry negative fitness effects (e.g., V_{max} decreasing mutations) along with beneficial escape mutations or reversions (genetic hitchhiking). Additionally, mutants with a small V_{max} increasing effect, but that are otherwise equal to the wild-type, may have a long fixation time and can easily be out-competed by, e.g., escape mutants. This makes within-host V_{max} evolution more selectively neutral, and hence more sensible in our model. In future work, we aim to test if these speculations are valid, and whether a more detailed within-host fitness and selection model can unify within-host evolution and population-level adaptation.

4.4 Methods

Our full model is a two-level individual and discrete-event based simulation, based on the Sellke construction [202]. The Sellke construction generalizes the Gillespie algorithm, by allowing for non-exponentially distributed waiting times. We need this generalization to allow for realistic non-exponential distributions of the length of the

	Agent	Events
within-host level	virus	escape mutation, reversion
	disease phase	transition to next disease phase
between-host level	host	transmission, death

Table 4.2: Agents and events in the model.

asymptomatic phase, as estimated earlier [65]. In our simulation events occur at particular points in time, which determines the order of these events. If an event takes place, this may alter the state (e.g. the number of susceptible individuals, or the virulence) and this influences the moments and order at which future events take place. The model was coded in `C++` and analyzed using `Python` and `R`. The source code can be downloaded from `www.github.com/chvandorp/sellke-hiv-model`.

The agents and events that are described explicitly in our model are listed in Table 4.2. In order to determine what the next event will be and when it takes place, we need to know how to compute waiting times.

4.4.1 Waiting times

In general, whenever a new event E is created during the simulation at time t, the exact moment when E will take place is unknown. Therefore, we assign to E a threshold τ_E and a load $\alpha_E = 0$. The threshold τ_E is sampled from some probability distribution Ψ_E with non-negative support and mean 1. We first compute the waiting time w_E, while conditioning on E being the first event to take place:

$$w_E = \inf\left\{w : \tau_E = \alpha_E + \int_t^{t+w} \mu_E(s)\,\mathrm{d}s\right\}. \tag{4.3}$$

Here, μ_E is the rate (or hazard) at which E takes place, which can depend on time. Notice that the (conditional) waiting time w_E could be infinite (e.g., when the number of susceptibles equals zero, the first event to take place can never be a transmission).

When we perform this computation for all future events E, we find the event F that must take place first, and also the time at which it takes place, i.e., $t + w_F$. We then perform the following steps: First, we update the time $t \mapsto t + w_F$. Then, for all future events E, we update the load α_E as follows:

$$\alpha_E \mapsto \alpha_E + \int_t^{t+w_F} \mu_E(s)\,\mathrm{d}s\,. \tag{4.4}$$

E	Ψ_E	μ_E	actions
within-host level			
escape mutation	Exp(1)	$\lambda_{\text{esc}}(t) \cdot (k-e)$	$e \mapsto e+1$, create new escape mutation event
reversion	Exp(1)	$\lambda_{\text{rev}}(t) \cdot f$	$f \mapsto f-1$, create new reversion event
phase change ($i=1$)	$\mathbb{P}(\tau_E = 1) = 1$	D_1^{-1}	change the phase into "asymptomatic", create a new phase change event
phase change ($i=2$)	Gamma(ρ^{-1}, ρ)	$D_2(V(t))^{-1}$	change the phase into "AIDS phase", create a new phase change event
phase change ($i=3$)	$\mathbb{P}(\tau_E = 1) = 1$	D_3^{-1}	end of the within-host simulation
between-host level			
transmission	Exp(1)	$\beta(t) \cdot S/N$	$S \mapsto S-1$, create a new infected individual, create a new transmission event for both the transmitting and the receiving host
death	$\mathbb{P}(\tau_E = 1) = 1$	$(t_{\text{death}} - t_{\text{infect}})^{-1}$	$S \mapsto S+1$, remove the deceased host

Table 4.3: Threshold distributions, rates, and actions for the events in the model. The functions λ_x for $x \in \{\text{esc}, \text{rev}\}$ are here defined by $\lambda_x(t) = \lambda_{1,x}$, if the patients disease is in the acute phase, and similarly $\lambda_x(t) = \lambda_{2,x}$ for the asymptomatic phase, and $\lambda_x(t) = \lambda_{3,x}$ for the AIDS phase. The function $t \mapsto V(t)$ describes the viral load during the asymptomatic phase (that may not be constant due to escape mutations and reversions). The shape parameter $\rho = 3.46$ for the Gamma distribution was estimated by Fraser *et al.* [65].

Finally, we let the event F act on the current state and remove F from our event list. For instance, if F happens to be a transmission event, we should initiate a new host and decrease the number of susceptible individuals. Additionally, new transmission events should be created for the transmitter and recipient. Hereafter, we re-compute the waiting times w_E for all events E and repeat the above steps.

In most cases, the computation of w_E and the updating of α_E is simple. For instance, if E is a reversion event and $f > 0$, then $w_E = \frac{\tau_E - \alpha_E}{\lambda_{i,\mathrm{rev}} f}$. For updating the load, we replace α_E with $\alpha_E + w_F \lambda_{i,\mathrm{rev}} f$. A transmission event requires more effort, because the rate of transmission varies during an individual's infectious lifetime.

The model can now be described by specifying for the events E listed in Table 4.2, their threshold-distribution Ψ_E, their rates μ_E, and the precise actions on the state (see Table 4.3).

4.4.2 Stochastic computation of the next-generation matrix

In order to study the steady state of the above described model, we developed a faster and more accurate method. In deterministic (e.g. ODE-based) models with multiple viral strains, one can compute the next-generation matrix (NGM), using the model's rate equations [55]. Given a generation (i.e., a distribution of strains in a cohort of newly infected individuals), the NGM gives the next generation after mutation and selection in a discrete generation-based model. The steady states of the original (continuous time) and generation-based model coincide. This steady state can be computed by finding the dominant eigenvector of the NGM. The dominant eigenvalue equals (by definition) $\mathscr{R}_0$ [46].

Our model is not deterministic, but we can approximate the NGM using a Monte-Carlo method. We start with a virus that has m_1 mutations. We then infect a large cohort (of size N) of individuals. These individuals may have different binding repertoires (of diverse size k), so we first sample pairs (e_1, f_1) with $e_1 \sim \mathrm{Hyper}(m_1, k, n)$ and $f_1 = m_1 - e_1$. Then we run a within-host simulation for each of the virus-host pairs. Finally, we sample strains (e, f) that would be transmitted by the hosts at the start of an epidemic, and we count the number of transmitted stains C_{m,m_1} that have $m = e + f$ mutations. The vector $(\hat{G}_{m,m_1})_{m=0,\ldots,n}$ with $\hat{G}_{m,m_1} := C_{m,m_1}/N$ approximates the m_1-th column of the NGM.

If the sample size N is large enough, the dominant eigenvalue and corresponding right eigenvector of the matrix $\hat{G} = (\hat{G}_{m,m_1})_{m,m_1=0,\ldots,n}$ approximate, respectively, $\mathscr{R}_0$ and the steady state distribution of prevalent viral strains in our agent-based model. By sampling strains from the steady-state distribution, and simulating infections with these strains, we can compute statistics as $\langle \mathrm{spVL} \rangle$ in equilibrium. This method is not

based on formal arguments, but below we put forward some heuristic evidence for its correctness.

4.4.3 Estimating heritability

For the statistic heritability (h^2), the above scheme is insufficient. However, we do have a cohort of potential transmitters, and hence we can create transmission couples by first sampling transmitted strains from the cohort's individuals, and then infecting recipients. The statistic h^2 is computed as the slope of the regression between the spVL of transmitters and receivers.

Classically, heritability of a trait x is defined as the proportion of variance in x that is caused by inherited genetic factors [see e.g. 206]. Hence, if we write $x = \gamma + \varepsilon$, where γ is a genetic, and ε an environmental factor, then $h^2 := \mathrm{Var}(\gamma)/\mathrm{Var}(x)$. The slope of the regression mentioned above is an estimator for this quantity, but only when the transmitted quantity $\gamma' = \gamma +$ "mutational error" in the recipient is independent of the the transmitter's environmental factor ε. Below we will see that such an independence assumption does not hold for our model, and that the use of transmission couples results in an over-estimate of spVL heritability.

4.4.4 Quantification of the footprint effect on heritability of spVL

To quantify the effect of the immunological footprint on heritability, we use a structural equation model (SEM), depicted as a directed, acyclic graph (DAG) in Figure 4.7. In our model, the actual inherited quantity is the number of mutated peptides $m = e + f$. During an infection this quantity can change due to escapes and reversions, so we will only incorporate the number of mutations at the moment of infection (m_{tra} for a transmitting host, and m_{rec} for the corresponding receiver) in our statistical model.

The set-point virus load of the receiver ($\mathrm{spVL}_{\mathrm{rec}}$) depends on m_{rec}, and the breadth of the immune response against the wild-type virus (k_{rec}). Of course, more factors determine the set-point virus load, such as the initial number of escape mutations, and stochastic effects such as mutations and progression to AIDS, but the simplified SEM only contains the variables spVL, m and k.

Apart from k_{tra}, the breadth of the transmitter's immune response and m_{tra}, the set-point virus load of the transmitter ($\mathrm{spVL}_{\mathrm{tra}}$) depends also on m_{rec}. This is because the set-point is an average over the chronic phase, and hence, the transmitted virus co-determines the set-point of the transmitter. In Figure 4.7, this is indicated by the arrow $m_{\mathrm{rec}} \longrightarrow \mathrm{spVL}_{\mathrm{tra}}$.

During infection of the transmitter, the virus escapes a number of immune responses, and this number is dependent on k_{tra}. This means that k_{tra} influences the number of mutations of the transmitted virus m_{rec}. This immunological footprint is represented by the arrow $k_{\text{tra}} \longrightarrow m_{\text{rec}}$ in Figure 4.7. The breadth of the immune response k_{tra} has no direct effect on m_{tra}, since m_{tra} corresponds to the transmitter's founder virus. Likewise, there is no direct effect of k_{rec} on m_{rec}.

We use the the `R` package `lavaan` [193] to fit the model to (standardized) simulated data, that were produced using the NGM method and the standard model's parameters. As an example, the result of one of such fits is given in Figure 4.7. In this graph, the numbers above the arrows indicate the estimated weights. The maximal virus load V_{max} equals $6.64\log_{10}$ copies per ml, and the mutation rate $\lambda_{1,\text{esc}}$ equals 3y^{-1}, such that the mean set-point for this population is $4.51\log_{10}\text{ml}^{-1}$ (cf. Figure 4.4A). Despite the large sample size of 25 690 transmission couples, and the fact that the SEM has 4 degrees of freedom, the model describes the data quite well (the χ^2-test's p-value equals 0.81, and the root mean square error of approximation (RMSEA) equals 0 with a 90% CI of $[0, 0.01]$).

In the context of our SEM, the statistic h^2 equals the correlation between spVL_{rec} and spVL_{tra}. This correlation can be computed as the sum of the contribution of all paths that connect spVL_{tra} with spVL_{rec}. The contribution of each path equals the product of the coefficients along the path. The 3 paths that connect spVL_{tra} with spVL_{rec} are:

$$\begin{aligned} P_1 &:= \text{spVL}_{\text{tra}} \longleftarrow m_{\text{rec}} \longrightarrow \text{spVL}_{\text{rec}}\,, \\ P_2 &:= \text{spVL}_{\text{tra}} \longleftarrow m_{\text{tra}} \longrightarrow m_{\text{rec}} \longrightarrow \text{spVL}_{\text{rec}} \quad \text{and}, \\ P_3 &:= \text{spVL}_{\text{tra}} \longleftarrow k_{\text{tra}} \longrightarrow m_{\text{rec}} \longrightarrow \text{spVL}_{\text{rec}}\,, \end{aligned} \tag{4.5}$$

where P_3 is responsible for the immunological footprint. In the example of Figure 4.7, the contribution of P_3 equals 0.082, which is about half (49.7%) of the total correlation between spVL_{rec} and spVL_{tra} (i.e., of the heritability). We refer to the contribution of the path P_3 as the "contribution of the immunological footprint to heritability".

4.4.5 A comparison of HIV-1 clades B and C

We downloaded representative sequences for clades B and C from LANL's HIV sequence database (`www.hiv.lanl.gov`; four sequences for each clade, as described by Leitner *et al.* [121]). Then, we downloaded the HLA-A and HLA-B distributions for Europe and Sub-Saharan Africa from the NCBI database dbMHC [`www.ncbi.nlm.nih.gov/projects/gv/mhc`; 151]. Using the MHC binding predictor NetMHCpan [version 2.4; 99], we computed binding affinities of all 9-mers from the representative strains for the most common HLA alleles (covering 95% of the populations). For each HLA molecule,

the binding threshold was chosen such that the top 1% of a set of 10^5 naturally occurring peptides would be considered a binder (as described by van Deutekom *et al.* [43]).

For our analysis, we sample pairs of HLA haplotypes from the HLA distributions of one of the populations (ignoring linkage disequilibrium), each haplotype consisting of two HLA-A alleles and two HLA-B alleles. For each two haplotypes, we then compare the similarity of the binding repertoires with respect to one of the four representative strains. As a measure of similarity, we use the Jaccard index (J): the size of the intersection, divided by the size of the union of the two binding repertoires. This gives us the distribution of similarity scores of a population with respect to a strain. Figure 4.8B depicts two of these distributions. The black bars correspond to the European population with respect to a clade B virus, and the gray bars to the Sub-Saharan population with respect to a clade C virus.

By comparing the similarity distributions of a Sub-Saharan with a European population (Figure 4.8B), we can assess the difference in heterogeneity between the two populations and clades. The right panel of Figure 4.8C depicts the medians (one value for each representative strain). The European medians are significantly higher than the Sub-Saharan medians. For a better comparison between two distributions, we use a U-statistic, defined as $U := \mathbb{P}(J_{\text{eur}} > J_{\text{afr}})$, where J_{eur} and J_{afr} are distributed as the European and Sub-Saharan similarity distributions, respectively (cf. the Mann-Whitney U-test). Hence, U equals the likelihood that a random haplotype pair in the European population shows more similarity than a random pair in the Sub-Saharan population. We have four clade B strains and four clade C strains, and hence we can compute 16 probabilities U (Figure 4.8C, right panel). They turn out to be significantly higher than 0.5, meaning that the European population, subject to clade B strains, is less heterogeneous than the Sub-Saharan population and clade C strains.

4.4.6 Deterministic computation of the NGM

We model within-host escape and reversion with two Markov chains:

$$e \xrightarrow{(k-e)\cdot\lambda_{i,\text{esc}}} e+1\,, \quad f \xrightarrow{f\cdot\lambda_{i,\text{rev}}} f-1\,. \tag{4.6}$$

Let $P_{i,t}(e|e_i)$ and $Q_{i,t}(f\,|f_i)$ denote the probability at time t that during infection phase i the host is infected by a virus with e escape mutations and f deleterious mutations, respectively, given that phase i started with an (e_i, f_i)-virus at time $t = 0$. These

probabilities satisfy the Kolmogorov forward equations [see e.g. 169].

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}P_{i,t}(e|e_i) &= \lambda_{i,\mathrm{esc}}(k-e+1)P_{i,t}(e-1|e_i) - \lambda_{i,\mathrm{esc}}eP_{i,t}(e|e_i)\\ \frac{\mathrm{d}}{\mathrm{d}t}Q_{i,t}(f|f_i) &= \lambda_{i,\mathrm{rev}}(f+1)Q_{i,t}(f+1|f_i) - \lambda_{i,\mathrm{rev}}fQ_{i,t}(f|f_i)\,.\end{aligned} \tag{4.7}$$

Closed-form expressions for $P_{i,t}(e|e_i)$ and $Q_{i,t}(e|e_i)$ are given by

$$\begin{aligned}P_{i,t}(e|e_i) &= \binom{k-e_i}{k-e}\exp\left(-\lambda_{i,\mathrm{esc}}t\right)^{k-e}\left(1-\exp(-\lambda_{i,\mathrm{esc}}t)\right)^{e-e_i}\\ Q_{i,t}(f|f_i) &= \binom{f_i}{f}\exp\left(-\lambda_{i,\mathrm{rev}}t\right)^{f}\left(1-\exp(-\lambda_{i,\mathrm{rev}}t)\right)^{f_i-f}\,.\end{aligned} \tag{4.8}$$

The probability $P_{i,t}(e|e_i)\cdot Q_{i,t}(f|f_i)$ that the host is infected with an (e,f)-virus only makes sense if we condition on the infection still being in phase i. We want to get the expected number of transmitted virus of a specific type, and in order to make the calculations possible, we take exponential distributions for the length of the phases. We tested that this assumption is not crucial by also considering Erlang distributions. The rate at which phase i ends is given by $\delta_i = 1/D_i$. We also assume that mutation during the asymptomatic phase is slow and that the spVL is determined by the virus at the end of the acute phase (which is of type (e_2,f_2)). This means that β_2 and δ_2 can be kept constant. Furthermore, the fraction of susceptible individuals ($s := S/N$) can be kept constant, either because the population is in a steady state, or because the epidemic has just started ($s \approx 1$).

Consider the probability generating function [cf. 45] for the number of transmitted virus of type (e,f) during phase i:

$$g_i(e,f|e_i,f_i;z) := \int_0^\infty P_{i,t}(e|e_i)Q_{i,t}(f|f_i)\delta_i\exp(-\delta_i t + (z-1)\beta_i s t)\,\mathrm{d}t\,. \tag{4.9}$$

Assuming that $\lambda_{i,\mathrm{esc}} = \lambda_{i,\mathrm{rev}}$, we can write this integral in terms of the Beta function (B). First we substitute the above given expressions for P and Q

$$\begin{aligned}g_i(e,f|e_i,f_i;z) = \delta_i\binom{k-e_i}{k-e}\binom{f_i}{f}\int_0^\infty &\exp(-\lambda_{i,\mathrm{esc}}(k-e_i)t - \lambda_{i,\mathrm{rev}}f_i t + (z-1)\beta_i s t)\\ &\times(1-\exp(-\lambda_{i,\mathrm{esc}}t))^{e_i-e}(1-\exp(-\lambda_{i,\mathrm{rev}}t))^{f-f_i}\,\mathrm{d}t\end{aligned} \tag{4.10}$$

and when we now assume that $\lambda_{i,\text{esc}} = \lambda_{i,\text{rev}}$, we can get

$$\begin{aligned} g_i(e,f\,|e_i,f_i;z) &= \delta_i \binom{k-e_i}{k-e}\binom{f_i}{f} \int_0^\infty \exp(-\lambda_{i,\text{esc}}(k-e_i+f_i)t + (z-1)\beta_i s t - \delta_i t) \\ &\quad \times (1-\exp(-\lambda_{i,\text{esc}}t))^{e_i-e+f-f_i}\,\mathrm{d}t \\ &= \frac{\delta_i}{\lambda_{i,\text{esc}}}\binom{k-e_i}{k-e}\binom{f_i}{f} \int_0^1 u^{k-e_i+f_i-\frac{(z-1)\beta_i s-\delta_i}{\lambda_{i,\text{esc}}}-1}(1-u)^{e_i-e+f-f_i}\,\mathrm{d}u \end{aligned} \tag{4.11}$$

which equals by definition

$$\frac{\delta_i}{\lambda_{i,\text{esc}}}\binom{k-e_i}{k-e}\binom{f_i}{f} B\left(k-e_i+f_i-\frac{(z-1)\beta_i s-\delta_i}{\lambda_{i,\text{esc}}}, e_i-e+f-f_i+1\right). \tag{4.12}$$

Since one of the arguments in this Beta function is an integer, the function g_i is rational:

$$g_i(e,f\,|e_i,f_i;z) = \frac{\delta_i}{\lambda_{i,\text{esc}}}\binom{k-e_i}{k-e}\binom{f_i}{f} \frac{(e-e_i+f_i-f)!}{\left(k-e+f-\frac{(z-1)\beta_i s-\delta_i}{\lambda_{i,\text{esc}}}\right)_{e-e_i+f_i-f+1}}, \tag{4.13}$$

where we use the (rising) Pochhammer symbol $(x)_r := x(x+1)\cdots(x+r-1)$.

Now that we have this expression for g_i, we can exploit the probability generating function's useful properties. The number $g_i(e,f\,|e_i,f_i;1)$ equals the probability that at the end of phase i, the host is infected with an (e,f)-virus. The expected number of transmitted (e,f)-strains during phase i equals $\frac{\partial}{\partial z}\big|_{z=1} g_i(e,f\,|e_i,f_i;z)$. We use the following notation:

$$\begin{aligned} \psi_i(e,f\,|e_i,f_i) &:= \tfrac{\partial}{\partial z}\big|_{z=1} g_i(e,f\,|e_i,f_i;z) \\ \pi_i(e,f\,|e_i,f_i) &:= g_i(e,f\,|e_i,f_i;1) \end{aligned} \tag{4.14}$$

If we now take into account that a transmitted (e,f)-virus has a different phenotype (e',f') in the receiver (with probability given by the hypergeometric distribution), we can find the NGM for the case $\lambda_{i,\text{esc}} = \lambda_{i,\text{rev}}$. We verified that for this part of the parameter space (i.e., $\lambda_{i,\text{esc}} = \lambda_{i,\text{rev}}$), the deterministic and stochastic computation give the same results (not shown).

4.4.7 The bifurcation in the model with a homogeneous host population

When the host population is homogeneous ($n = k$), we find a threshold in the parameter space across which between-host adaptation is no longer possible. Here we will make this precise and show that this threshold is caused by a transcritical bifurcation. In a homogeneous population, we lose deleterious mutations. In the notation introduced above, we may ignore $f = 0$ and we write for instance $\pi_i(e|e_i) := \pi_i(e,0|e_i,0)$. Let $G = (G_{e,e_1})_{e,e_1}$ denote the NGM, then we get the following formula in terms of π_i and ψ_i:

$$G_{e,e_1} = \psi_1(e|e_1) + \sum_{e_2=0}^{k} \pi_1(e_2|e_1)\left(\psi_2(e|e_2) + \sum_{e_3=0}^{k} \pi_2(e_3|e_2)\psi_3(e|e_3)\right) \tag{4.15}$$

The matrix G is triangular, since the number of escape mutations, which equals the total number of mutations, can only grow during an infection. The diagonal elements of G are the eigenvalues of G, and the dominant eigenvalue equals (by definition) $\mathscr{R}_0$ of the quasi-species. The diagonal elements can be written as

$$G_{\ell,\ell} = \psi_1(\ell|\ell) + \pi_1(\ell|\ell)(\psi_2(\ell|\ell) + \pi_2(\ell|\ell)\psi_3(\ell|\ell)), \quad \ell = 0,\dots,k. \tag{4.16}$$

If $G_{k,k}$ is dominant, then population-level evolution will result in strains that have escaped all CTL responses. If another eigenvalue $G_{\ell,\ell}$ with $\ell < k$ is dominant, then not all viruses have escaped all CTL responses and this is due to selection for transmission on the population-level.

If we now fix $V_{\max}$ and let $\lambda_{1,\text{esc}}$ approach 0 from the right, then for high $\lambda_{i,\text{esc}}$ the eigenvalue $G_{k,k}$ is dominant. The mentioned bifurcation occurs when $G_{k,k}$ equals one of the $G_{\ell,\ell}$ (with $\ell < k$) for the first time. We first give simple expressions for π_i and ψ_i that occur in the expression for $G_{\ell,\ell}$:

$$\begin{aligned} \pi_i(\ell|\ell) &= \frac{\delta_i}{(k-\ell)\lambda_{i,\text{esc}} + \delta_i} \\ \psi_i(\ell|\ell) &= \frac{\delta_i \beta_i s}{((k-\ell)\lambda_{i,\text{esc}} + \delta_i)^2}. \end{aligned} \tag{4.17}$$

These expressions and the formula for $G_{\ell,\ell}$ enable us to (numerically) find the curves $\{G_{\ell,\ell} = G_{k,k}\}$ for $\ell = 0,\dots,k-1$. These curves and the resulting threshold are shown in Figure 4.6.

Chapter 5

Modeling immunological pre-adaptation of HIV-1

Christiaan H. van Dorp[1, 2], Michiel van Boven[2], and Rob J. de Boer[1]

[1]Theoretical Biology and Bioinformatics, Universiteit Utrecht, Utrecht, The Netherlands.
[2]National Institute for Public Health and the Environment, Bilthoven, The Netherlands.

In preparation

Abstract. It is becoming increasingly evident that the evolution of HIV-1 is to a large extent determined by the immunological background of the host. On the population-level this results in associations between specific human leukocyte antigen (HLA) alleles and polymorphic loci of the virus. Furthermore, some HLA alleles that were previously associated with slow progression to AIDS have been shown to lose their protective effect, because HLA-specific immunological escape variants spread through the population. This phenomenon is known as immunological pre-adaptation. Apart from adapting to human immune responses, the set-point virus load (SPVL) of HIV-1 is thought to have evolved to values that optimize the population-level fitness of the virus. This suggestion is supported by considerable heritability of the SPVL. Previous modeling studies show that whether or not SPVL optimization is expected to occur depends sensitively on the underlying assumptions with respect to the extent of within- versus between-host selection. Here we use a detailed and semi-realistic multi-level HIV-1 model in which immunological pre-adaptation and SPVL evolution can emerge from the underlying interactions of the virus with the immune system of the host. This enables us to study the effect of immunological escape on disease progression, and how disease progression may be molded by SPVL evolution. We find that the time to AIDS could decrease significantly (0.5-1.0 years) in a HLA-dependent manner by immunological pre-adaptation over the long-term course of the epidemic (> 100 years). We find that SPVL is not expected to evolve to optimize the population-level fitness of HIV-1, even though high heritability of the SPVL emerges from continual selection of immune-escape mutations.

5.1 Introduction

Human immunodeficiency virus type 1 (HIV-1) virulence evolution has been studied extensively. High virulence is related to fast disease progression, but also to high set-point virus load (SPVL) and increased infectiousness. A corollary is a life-history trade-off between infectiousness and longevity, which has been hypothesized to steer the evolution of virulence and SPVL [65]; a hypothesis for which there is mounting evidence [19, 22]. Complicating the idea is the fact that selection operates at different levels [66, 93], and potentially shapes HIV-1's fitness landscape in a non-trivial manner. Recently, the characteristics of this fitness landscape itself have become a subject of study [10, 59, 94, 113, 132]. In addition, the environment (i.e. the host) impacts the shape of the fitness landscape, as evidenced by strong virus-host interactions [146].

Due to genetic variability of the host, notably HLA-polymorphism, the fitness of HIV-1 is not an invariant global property. Instead, after every transmission event new immune responses are to be escaped, while escapes from the transmitter's cytotoxic T lyphocytes (CTLs) may be reverted whenever deleterious. Not all deleterious mutations are reverted after transmission, as is clear from analyses of HLA footprints on viral populations world-wide [106]. In these cases, compensatory mutations can still (partially) restore viral fitness. Further, many of these mutations have epistatic effects, and many epistatic interactions are in fact host-mediated [26]. This process of escape, reversion and compensation drives the evolution of HIV-1 over a complex and apparently rugged fitness landscape.

What is meant by fitness in earlier studies depends on the context. For instance, for *in-vitro* experiments, the virus replicative capacity (vRC; the rate of exponential growth) is often used as a fitness measure (the Malthusian fitness). On the population level, the expected number of secondary cases caused by one infected individual is thought to be a good notion of fitness, and is equal to the basic reproduction number ($\mathscr{R}_0$) at the onset of the epidemic. In each case, fitness is the product of viral and host traits. Non-linear and highly stochastic within-host processes determine how individual-level fitness is related to population-level fitness. Here, we model such processes in detail, to gain insight in how the levels of selection are connected.

A number of multi-level HIV-1 models have been developed before. Notably, Lythgoe *et al.* [139] experiment with rugged fitness landscapes, and a dynamic within-host model. Doekes *et al.* [49] extended this model with a reservoir of latently infected cells. Hool *et al.* [100] use a within-host model of HIV-1 to determine the within-host evolutionary dynamics, and population-level parameters. These studies try to reconcile within-host evolution and selection with the adaptive character of SPVL evolution. Previously, we have modeled HIV-1 virulence evolution in a HLA-polymorphic population [54], and found that the SPVL distribution only obeys transmission-adaptation when

escape and reversion rates are unreasonably low. Some modeling choices in the previous study were influenced by computational convenience. Here, we follow a bottom-up approach that addresses the more *ad hoc* simplifications made before. Specifically, the within-host processes are modeled in detail, and the higher level variables are derived from the within-host model.

In order to couple a between-host with a within-host model, the infectiousness and the duration of the infectious period need to be derived from within-host processes. For the infectiousness and duration we use the empirical relation with virus load [19, 65]. Hence, we require a model that yields viral load from the within-host dynamics. Our analyses are based on a widely used system of ordinary differential equations [5, 44], describing the relation between target cells ($CD4^+$ cells), HIV-1 infected cells, and effector cells (HIV-1 specific CTLs). Apart from the virus load, this model provides us with rules for the introduction of new mutants and the addition of new immune responses. The model allows for clonal interference and slow rates of immunological escape. Both phenomena have been described for CTL escapes [172, 188].

Separately, the key ingredients of our HIV-1 immuno-epidemiological model can be found in the three examples of nested models mentioned before [54, 100, 139]. The model presented here combines all of these components: (i) The parameters determining the viral dynamics are determined by a complex and rugged genotype-to-phenotype mapping, as was touched upon by Lythgoe *et al.* [139]. We improve upon this by using the *NK* model [105] that has been developed as an abstract fitness model with epistatic interactions, and by allowing multiple parameters to be determined by the virus, instead of just the replicative capacity. (ii) The virus load, the resulting epidemiological parameters, and individual-level growth rates are calculated using a realistic model of within-host HIV-1 dynamics, akin to the work of Hool *et al.* [100], which is improved upon by allowing for co-existence and competition of multiple viral strains. (iii) Individual-level evolution is directed, but because immune responses are restricted to polymorphic HLA alleles, the effects of a particular mutation can be different in each host. We also implemented this aspect previously [54, Chapter 4]. Here, we further add the possibility of compensatory mutations (through the *NK* model), HLA alleles with life-like population-level frequencies, and immuno-dominant and sub-dominant immune responses.

Interest in population-level adaptation of HIV-1 is certainly not restricted to the SPVL. It has been known for a long time that certain HLA molecules are associated with a low SPVL and slow disease progression [108, 123]. For instance the HLA-B*57 and HLA-B*27 types are over-represented in so-called elite controllers: patients with a very low virus load in the absence of antiretroviral therapy. In more recent years, advanced statistical methods have been developed to discover associations between HLA molecules and viral polymorphisms [28], revealing widespread footprints of HLA in the genome of HIV-1.

A consequence of widespread escape from HLA-restricted immune responses is that individuals can be infected with viruses that are already adapted to their immune responses, because the same HLA molecule was encountered in a previous host. This phenomenon is referred to as "pre-adaptation". Pre-adaptation can lead to the apparent loss of HLA associations, if it is not corrected for [27]. Furthermore, evidence is accumulating that HLA pre-adaptation is a major determinant of disease progression [153], although currently the population-level health effects appear to be relatively small [110].

The spread of CTL escape mutations through the human population has been modeled before, but these models are restricted to only a few epitopes and HLA alleles [68, 69]. In the model presented here, we can follow escape mutations on a much larger scale. Furthermore, our bottom-up modeling approach allows us to investigate SPVL evolution, heritability, and the detrimental effects of pre-adaptation in an integrated manner.

5.2 Results

5.2.1 A multi-level HIV-1 model

Following previous modeling studies [88, 89], we simulate a population of individuals that can infect each other via a dynamic sexual contact network (Figure S5.1). The rate at which a susceptible individual gets infected depends on the number of infected partners, and each partner's transmission rate. The transmission rate of an infected individual, in turn, depends on the current virus load (VL), which is determined by the individual-level model (see below). The relation between VL and transmission rate has been estimated in a cohort of transmission pairs from Zambia [65]. The VL also relates to the risk of progression to AIDS, for which we use estimates from a cohort from Amsterdam [65]. We do not explicitly model an AIDS phase, but directly remove individuals from the population after the asymptomatic phase of the infection.

At the individual level, we model populations of susceptible $CD4^+$ T cells (target cells), infected cells and $CD8^+$ T cells (effector cells). We do not explicitly model the B-cell response. The dynamics of these populations is given by a system of ordinary differential equations. New immune responses and mutant virus strains are added to the system at random points in time, determined by an antigen-dependent immune-activation rate, and the mutation rate of the virus, respectively. An example of an individual-level simulation is given in Figure 5.1.

In previous modeling studies, the within-host fitness of the virus is typically only modulated by one or two parameters of the model [e.g. 44, 136]. Large variation in the SPVL can be achieved by allowing many parameters to vary by a small amount [161].

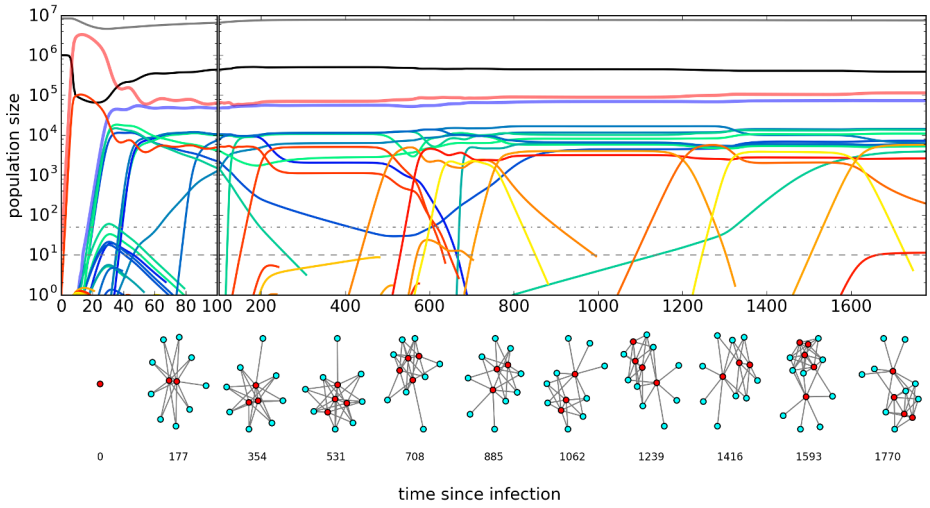


Figure 5.1: An example of an individual-level simulation. The top panel shows the time series of the population sizes of quiescent CD4$^+$T cells (gray), target cells (black), infected cells (yellow-red), and effector cells (green-blue). The thick lines indicate the total virus load (red) and the total number of effector cells (blue). The horizontal dashed line marks the population size at which declining populations are considered endangered (τ_{end}, see Methods), and the dot-dashed line pinpoints the population size at which we start tracking the mutants of a viral strain (τ_{mut}, see Methods). The networks indicate the interactions between immune responses (blue) and viruses (red). An immune response is connected with a virus if the virus contains the response's epitope. Two different viruses are connected if they differ at precisely one locus. Figure S5.2 shows the individual-level dynamics for different mutation rates, but otherwise identical initial conditions and parameters.

Here, we generalize this idea by putting multiple parameters of the ODE model under control of the virus. The genome of a virus is modeled using a sequence of $N = 540$ bits, and consists of 9 genes of equal length. The genes are responsible for specific functions of the virus. This is accomplished by letting each gene encode the value of a parameter in the system of ODEs. The default parameter values are listed in Table 5.1; the genes determine a deviation from the default (see Methods).

The most obvious parameter under control of the virus is the infection rate of target cells (β). The gene encoding β is analogous to the *env* gene that encodes the gp120 and gp41 proteins that mediate the attachment and entry to the target cell. After entry of the target cell, the infected cell is in an eclipse phase before it starts producing virions. The length of the eclipse phase (regulated by the parameter γ) is determined by another gene. This gene is analogous to the *pol* gene that encodes reverse transcriptase (RT) and integrase. The integration of the viral genome into the host cell's DNA has a high probability of failing, which is referred to as abortive infection, often leading to apoptosis or pyroptosis of the cell [40]. We model this with a virus-controlled parameter f, representing the fraction of successfully infected cells. The gene encoding f is partly analogous to *vif*, which disrupts the innate intra-cellular APOBEC3 response [184].

Next, we have parameters for production (p_V) and degradation (d_V) of virions. Although many genes are involved in the formation of virions, one can think of the *tat* gene that regulates transcription. Degradation of virions is modulated by the stability of structural proteins, for instance p17 and p24, encoded by the *gag* gene. Even in the absence of an immune response, an infected cell experiences a virus-induced cytopathic effect [40]. We model this using a virus-controlled parameter α, which is added to the death rate of virion producing cells. The gene encoding α is analogous to e.g. the *nef* gene. The same *nef* gene accounts for the down-regulation of MHC. We model the expression of MHC using a virus-controlled parameter m, which is used to regulate the killing rate of the infected cell by effector cells.

HIV-1 is known to modulate the activation of $CD4^+$ T-cells, making them permissive to infection. This principle has been used before in modeling studies [9, 100], and in our model we use a virus-controlled parameter ω to boost the activation of quiescent $CD4^+$ T-cells. It is not well known how exactly HIV-1 influences the activation of $CD4^+$ T cells, and although many candidate mechanisms exists [9], in our model, the parameter ω is determined by a single gene.

Further, the mutation rate (μ) is under viral control. The process of reverse transcription is highly error-prone and results on average in one mutated nucleotide per infected cell. However, to make the simulations computationally feasible, it is impossible to simply choose $\mu = 1/N$ as a default value of the mutation rate. Merely taking a smaller mutation rate (e.g. $\mu = 10^{-6}$) is computationally feasible, but ignores the fact that a

high mutation rate comes with a fitness cost, and would allow viruses to evolve very large mutation rates. We therefore let $N^* > N$ denote the length of a full genome, and think of the N modeled loci as a small subset of all loci. Mutations in the $N^* - N$ un-modeled loci are assumed to be highly deleterious, which we implement using the probability of faithful replication $(1-\mu)^{N^*} \approx e^{-N^*\mu} =: q$.

The relation between genotype and parameter value (phenotype) is modeled using the NK model [105]. The NK model has been used in many different settings, including immunological models [33], and allows for epistatic interactions between loci, which means that the effect of a mutation depends on the genetic background in which the mutation occurs. The exact implementation of the NK model is given in the Methods section. In addition to the parameter variation determined by the virus, each host has slightly different (10% standard deviation) individual-level parameters, to account for host-effects that are not due to the HLA genotype.

The virus- and host-dependent parameters each have a different effect on the SPVL. Figure 5.2 shows a cross-sectional sample from a simulated epidemic (see below). The SPVL is defined as the average virus load over the infectious period (see Figure 5.2A for individual time series). The gene with the largest effect on the SPVL modulates the β parameter, but also the genes responsible for MHC down-regulation (m), virion production (p_V), virion degradation (d_V), and abortive infection (f) are important factors. The length of the eclipse phase (γ), the mutation rate (μ) and target cell activation (ω) only have a limited effect on the SPVL. From the side of the host, the parameters that influence the SPVL the most are related to the number of target cells: the influx of $CD4^+$ T-cell (s) and the death rate of uninfected target cells (d_T), which is in agreement with previous model predictions [23]. Other important host-dependent parameters are related to the immune response [170]. Indeed, the maximum proliferation rate (p_E) and the death rate (d_E) of effector cells decrease and increase the SPVL, respectively.

It has been known for a long time that immuno-dominance plays an important role in control of HIV-1 [171]. Furthermore, modeling suggests that the subsequent emergence of immune responses could aid early escape [44]. We therefore include both phenomena in our model. The system of ODEs used in our model has the property that escape becomes exceedingly difficult when the immune response is broad, since the contribution of each single response to the total killing rate is small [44, 70]. It follows that equipping each host with too many responses in an early stage of the infection would retard viral escape and therefore individual-level evolution. On the other hand, limiting the number of responses leads to the unnatural situation where during many infections all responses are escaped. This problem is solved with immuno-dominance.

In the individual-level model, a number of CTL responses is in an *active* state, and these responses contribute to the killing rate. The remainder of potential responses can

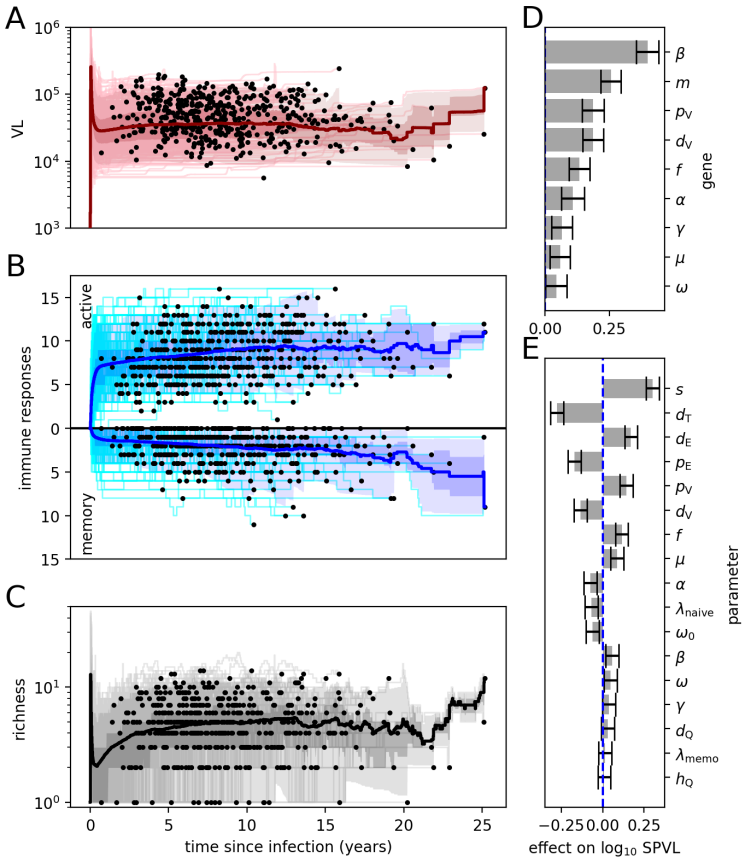


Figure 5.2: Individual-level time series from a cross-sectional sample. The sample (of size 500) is taken 200 years into an epidemic. The colored bands indicate the 2.5, 25, 75, and 97.5 percentiles, the thick line the average. The end of an infection is indicated with a black dot. **A.** Time series of the virus load (VL). **B.** Time series of the number of activated immune responses. The number of responses that are in the memory state (see Figure 5.9) are mirrored in the horizontal axis. **C.** Time series of the individual-level diversity, showing the number of strains (richness). **D.** The effect of the genes on the SPVL. A gene encodes an auxiliary parameter (w) using the NK genotype-to-phenotype map (see Methods). This w is used to scale the default value of a particular parameter in the system of ODEs. The effect on the standardized $\log_{10}$ SPVL of the w parameter of each gene is estimated with multiple linear regression. The error bar indicates the standard error or the estimate. **E.** The effect of the random host effects on the SPVL. The individual-level parameters are slightly different for each host. The effect of this variation on the standardized $\log_{10}$ SPVL is again estimated with multiple linear regression.

either be *trivial*, i.e. the virus does not express the epitope, or be in a *latent* state. The latent responses can become activated (at rate λ_{naive}) when e.g., an active response is escaped (Figure 5.9), and therefore the number of infected cells increases. When the antigen targeted by an escaped immune response re-appears, the escaped response can become re-activated. This happens at a faster rate than the activation of naive $CD8^+$ T cell ($\lambda_{\text{memo}} \gg \lambda_{\text{naive}}$).

In has been shown recently that after escape from one $CD8^+$ T-cell response, HIV-1 can induce *de-novo* $CD8^+$ T-cell responses [83]. In order to facilitate this behavior, each immune response targets a specific epitope, which is encoded by a locus in the genome and a particular combination of 3 bits. This is a simplification of a peptide epitope, usually 9 amino-acids long, for which the binding affinity to the HLA molecule is mostly determined by two anchor residues, and 6 other residues interact with the complement-determining region of the T-cell receptor (TCR). In our simplification, 1 bit determines HLA-binding, and 2 bits determine TCR-binding. In other words, the first bit determines if an immune response can be mounted against the epitope, and the remaining bits specify the response.

In a similar fashion to viruses determining parameters of the system of ODEs, immune responses control some of these parameters as well. Immuno-dominance is modeled by varying the amount of antigen needed for immune-response activation (h_{E}). The second parameter determined by the response is the killing rate (k). As we expect that immuno-dominant responses are also efficient at killing infected cells, the parameters h_{E} and k are negatively correlated.

The number of immune responses that is mounted varies between individuals and over the course of the infection (Figure 5.2B; active). A part of the activated responses are escaped or out-competed, and are de-activated (Figure 5.2B; memory). However, not every escape leads to subsequent de-activation of the immune response, as the virus population rapidly becomes diverse (Figure 5.2C), with each sub-population controlled by a different subset of the active immune responses (Figure 5.1; networks).

Although the number of target cells is not used to determine disease progression, ongoing escape from $CD8^+$ T-cell responses leads to a slow decline in the target cell count (Figure S5.3), which has been interpreted as the onset of AIDS in similar models [44]. In agreement with data [186], the slope of the $CD4^+$ T-cell count (ΔCD4) is negatively correlated with the SPVL (Figure S5.3B). Interestingly, the relation between ΔCD4 and SPVL appears to be non-linear, which has also been noticed by Regoes *et al.* [186]. A simple explanation for this non-linearity is given by the fact that ΔCD4 is naturally bounded from above: uninfected individuals and elite controllers have a $\Delta\text{CD4} \approx 0$.

5.2.2 Immunological escape generates genetic variation and heritability of the SPVL

When HIV-1 is transmitted to a new population, it encounters a set of new HLA molecules. To accommodate for this environmental change, we seed the simulated epidemic with a virus that is pre-evolved in a burn-in simulation. The burn-in simulation has identical parameters, except that the set of HLA molecules is completely different. The basic epidemiological and virological statistics are shown in Figure 5.3. The average vRC and SPVL are relatively stable. The SPVL slightly increases, presumably because of immune escape (see below). The SPVL is highly variable although the range is still two orders of magnitude lower than ranges observed in human populations. During the epidemic, the virus drifts away from the initial virus (wild type), while the genetic variation increases (Figure 5.3E). The increase in genetic variation coincides with an increase of the heritability of the SPVL, which reaches values around 20% (Figure 5.3D).

The relation between SPVL, infectiousness and the duration of the chronic phase results in a life-history trade-off. Hosts with an intermediate SPVL cause the most secondary infections. The fact that SPVL is heritable may result in the optimization of the basic reproduction number ($\mathscr{R}_0$) by means of SPVL evolution. The first evidence [65] that SPVL has evolved to optimize $\mathscr{R}_0$ comes from comparing the cross-sectional SPVL distribution with the shape of the transmission potential (TP; i.e. $\beta(V) \times D(V)$). Corroborating evidence comes from a longitudinal study [19], that shows that the mean SPVL changes over time, presumably towards an optimal value. The latter study suggests that the location of the transmission potential can differ between populations.

In order to test if similar behavior can be observed in our model, we conducted 5 simulations, starting with identical wild-type viruses and populations, but with slightly shifted TPs. The expectation is that the distribution of SPVLs approaches the shape of each TP. However, this expectation is not met in our model (Figure 5.4A), which is in line with the notion of short-sighted evolution [138, 139] and our previous work [54, Chapter 4]. Although optimization of the reproduction number is intuitively obvious, it is not necessarily the expected behavior from a mathematical perspective [128, 162]. We therefore added the mutation-at-transmission model [89, 206] to verify that under neutral individual-level evolution, $\mathscr{R}_0$-optimization is indeed expected to occur. In the 5 simulations with neutral SPVL evolution, the mean $\log_{10}$ SPVL evolves towards the $\mathscr{R}_0$-optimizing value (Figure 5.4B), confirming that our population-level model does not preclude $\mathscr{R}_0$-optimization.

Even though $\mathscr{R}_0$ is not optimized in our model, the virus is clearly subject to genetic drift as the genetic distance form the wild-type virus, as well as the genetic diversity in the population steadily increases as the epidemic progresses (Figure 5.3E). A closer look at the genetic diversity at each individual locus of the virus shows that polymorphisms

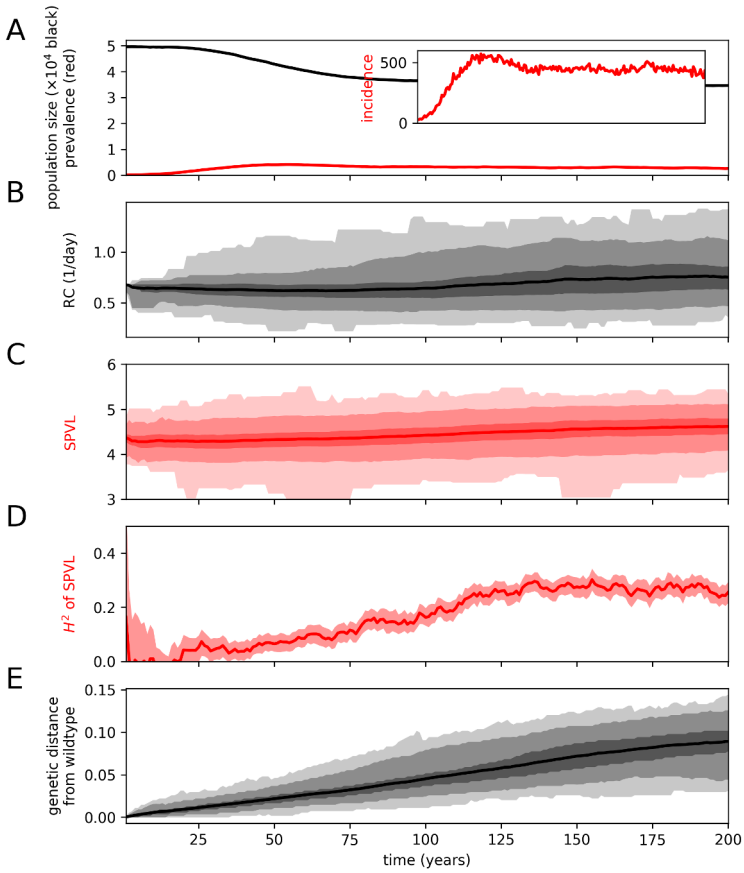


Figure 5.3: Statistics of the population-level simulation. A. Population size (black), prevalence of HIV-1 infected individuals (red) and incidence per year (inset) **B.** The generic replicative capacity (vRC) of the virus at the time of infection **C.** The set-point virus load (SPVL) **D.** The heritability (H^2) of the SPVL. **E.** The genetic distance from the wild-type virus (scaled by genome size). The diversity per locus is presented in Figure S5.4. The colored bands in panels **B**, **C** and **E** correspond to the 0, 2.5, 25, 75, 97.5 and 100th percentiles, and the 95% CI in panel **D**. The thick lines in panels **B**, **C**, **D** and **E** correspond to the mean values.

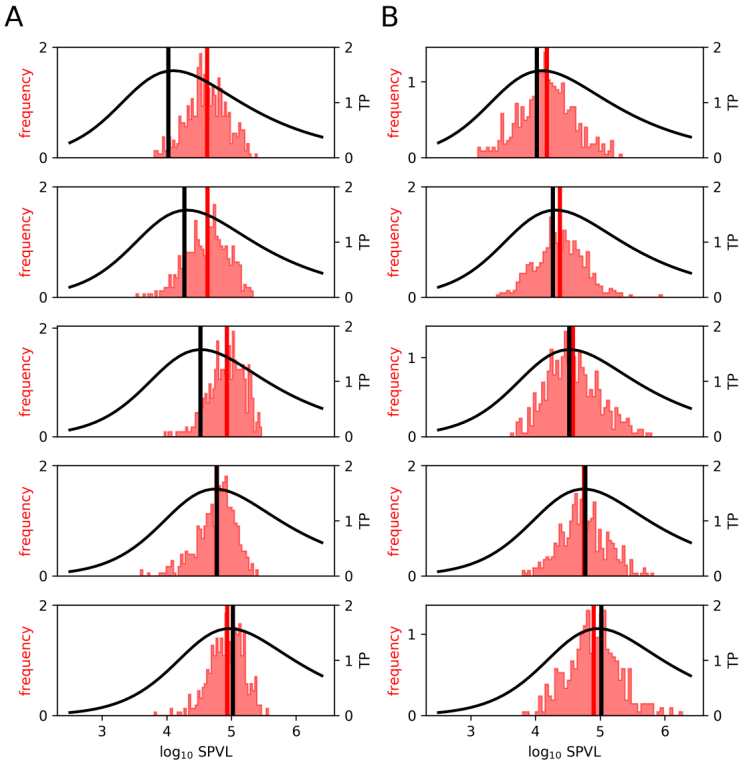


Figure 5.4: The SPVL distribution is insensitive to the location of the transmission potential. A. The 5 panels show the SPVL distribution after 300 years of evolution in identical populations (red histograms), except for the location of the transmission potential (TP; black curves). In each panel, the transmission rate β is shifted and scaled, resulting in transmission potentials with optima of identical height. However, these optima are attained at different SPVL values (black vertical lines). The population mean SPVLs are indicated by red vertical lines, **B.** The procedure is identical as in panel **A**, but in this case the mutation-at-transmission model from Shirreff *et al.* [206] was included (see Methods).

preferentially arise at loci that are targeted by frequent and immuno-dominant immune responses (Figure S5.4D). At the same time, loci at which mutation leads to a high fitness cost (as measured by vRC) are generally more conserved (Figure S5.4E). Hence, selection pressure from $CD8^+$ T-cell responses leads to a genetically diverse virus population at the epidemic level. This genetic variability impacts the individual-level parameters, which results in a positive heritability of the SPVL (Figure 5.3D).

5.2.3 Immunological pre-adaptation reshapes the HLA-dependent time to AIDS

HIV-1 polymorphisms are known to be associated with specific HLA alleles, which is a result of the adaptation of HIV-1 to $CD8^+$ T-cell responses [154]. These associations are lost over time as escape mutations spread through the population [122]. By correcting for the phylogenetic structure of the viral population, these lost associations can be recovered [26, 28, 68]. In order to see if this form of adaptation occurs in our model, we calculate every 5 years the odds ratios of observing viral loci that are mutated away from the wild-type virus given the presence of a particular HLA allele (see Methods). All significant associations are shown in Figure 5.5A. Many of these associations are only short lived, and are only observable during the first 50-100 years of the epidemic. However, when we correct for which allele was transmitted to the host (conveniently, we do not have to reconstruct a phylogeny to infer the transmitted allele) many of the short-lived associations become long-lived (Figure 5.5B).

The spread of CTL-escape mutations through the population can have a significant impact on the virulence, in a way that is dependent on the HLA genotype of the patient [27, 39, 110, 173]. Having full knowledge of the transmitter-founder virus, the exact genetic makeup of the virus that seeded the epidemic (the wild type), and all possible responses that a host is able to mount, it is relatively easy to quantify the impact of pre-adaptation on progression to AIDS in our model. Every 20 years, we make a clone of each infected individual (with identical HLA alleles and host-specific parameters), and infect this clone with the wild-type virus. We define the impact of pre-adaptation as the difference between the lengths of the infections of the clone and the original (Figure 5.6A; Δ length). To quantify pre-adaptation, we enumerate all possible immune responses against the transmitted virus, and compare this to all possible immune responses the clone can mount to the wild-type virus. The pre-adaptation is defined as the relative difference between the number of possible immune responses (weighted by immuno-dominance). This gives the extent to which the transmitted virus is pre-adapted to the receiving host as compared to the wild type.

As the epidemic progresses, the virus escapes from common and immuno-dominant immune responses, which coincides with a faster progression to AIDS (Figure 5.6A). The

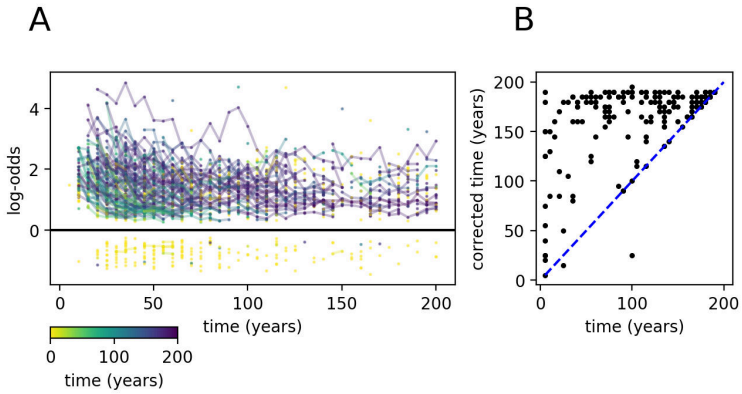


Figure 5.5: Associations between HLA alleles and viral polymorphisms disappear. A. Time series of the strength (log-odds) of associations between HLA alleles and viral polymorphisms during a simulated epidemic. The color of each time series corresponds to the length of the time interval in which the association is called significant. **B.** The time that an associations remains significant increases after correction for the transmitted virus. The length of the time interval in which an association was called significant is shown without (x-axis) and with (y-axis) this correction.

relation between pre-adaptation and increase in virulence becomes more pronounced as both become more variable (Figure 5.6B). The over-all effect of pre-adaptation is shown in Figure 5.6C, and is relatively small: the progression to AIDS is 0.5-1 year faster after 200 years of evolution. This is because most individuals are infected with a virus that is $\approx$ 30% pre-adapted. These results are roughly comparable with real data: On the one hand, the effect of being infected with a pre-adapted virus can be quite large [153]. However, the fraction of infections with highly pre-adapted viruses currently remains low [110]. A warning from this study would be that the level of pre-adaptation is expected to increase during the next century.

5.3 Discussion

We have developed a multi-level model of HIV-1 that incorporates a number of realistic attributes. The most notable features are: (i) A virus is represented by a genome that encodes multiple parameters of the within-host dynamics. (ii) The individual-level model allows for competition between multiple strains and escape from multiple immune responses that vary in immuno-dominance. (iii) The host population is highly heterogeneous, but individuals can share immune responses due to pre-defined HLA alleles. The model allowed us to study population-level evolutionary aspects

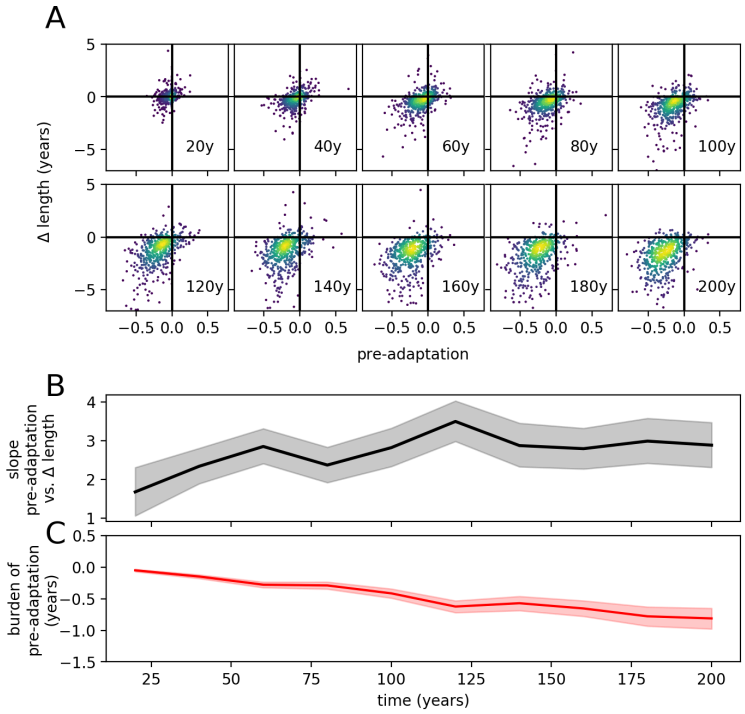


Figure 5.6: The burden of immunological pre-adaptation increases during the epidemic. **A.** The scatter plots show the relation between relative pre-adaptation of the TF virus with respect to the WT virus, and the difference in the duration of the infection before progression to AIDS. The plots were made at regular time intervals. The color indicates the density of the observations. **B.** At each time point, a slope is calculated with linear regression, which are displayed in the middle panel. The slope indicates the impact of a unit of pre-adaptation in terms of years lost. The band indicates the 95% CI. **C.** The bottom panel shows the average impact of pre-adaptation. This value (and the 95% CI) is calculated by multiplying the average pre-adaptation with the slope of the linear regression.

of HIV-1: both the evolution of the SPVL, and the adaptation of HIV-1 to human $CD8^+$ T-cell responses.

The model behaves realistically when it comes to the spread of escape mutations. Associations between HLA alleles and viral polymorphism arise quickly at the start of the epidemic, and subsequently disappear due to pre-adaptation. Infections with pre-adapted virus result in faster disease progression, and we predict that this impact of pre-adaptation will continue to grow.

Previously, we argued that because of short-sighted evolution within the host, SPVL evolution towards $\mathscr{R}_0$-optimizing values is not to be expected when the mutation rate is high. This was true even in a population with massive HLA polymorphism [54, Chapter 4]. However, we left open the possibility that $\mathscr{R}_0$-optimization could still emerge in a model with a more intricate fitness landscape that allows for within-host competition between strains. Here we have tested this idea, and have to conclude that our original expectation remains valid.

Although our within-host model is already relatively complex, it still misses a number of important aspects of HIV-1 dynamics. For instance, we do not explicitly model antibody responses, even though the antibody response might be a key factor in the breakdown of control of the virus [234].

Another ignored hallmark of HIV-1 is recombination. Recombination has been included in models that study, e.g., the decrease of the immune-escape rate with the time since infection [71, 72], as a result of clonal interference. Efficient recombination reduces the effect of clonal interference, thereby increasing the rate of individual-level adaptation by means of escape, reversion, and compensation [163], and hence could be of importance in our model. Furthermore, combining recombination with co-infection of the simulated hosts would result in a model that allows one to study the emergence of recombinant HIV-1 types.

An important obstacle to curing HIV-1 is the latent reservoir: a population of infected, but quiescent $CD4^+$ T cells that are invisible to the immune system and insensitive to antiretroviral drugs. This reservoir could be key to resolving short-sighted evolution with the optimization of $\mathscr{R}_0$ [138]. This idea has recently been successfully explored by Doekes *et al.* [49] with the aid of a nested multi-strain model. Their model gave the best results when ancestral strains were preferentially transmitted, which in turn can be understood in light of a selection bias at the transmission bottleneck [30], and would explain the discrepancy between within-host and between-host evolutionary rates of HIV-1 [1, 137].

Constructing a realistic within-host model of HIV-1 using ODEs that simultaneously allows for multiple orders of magnitude of SPVL variation between patients, as well as a high heritability of this SPVL has turned out to be quite a challenge. The many-genes method [cf. 161] gives reasonable results, but our SPVL range is still 2 orders of

magnitude narrower than realistic ranges. In part, this is because various biological mechanisms are not incorporated in our model. As an example, including HIV-1 tropism and hosts with the CCR5-Δ32 allele could account for some of the lower SPVLs, thereby widening the SPVL distribution. Alternatively, the "shifting-mosaic metapopulation" model [136] could be used to get better results, but incorporating mutation and competition in such a model will inevitably be challenging.

Finally, in spite of our strive for realism, we have used a relation between virus load and the duration of the infection that is still phenomenological. In a truly mechanistic model, the risk of AIDS would increase as the $CD4^+$ T-cell count gets close to 200 cells per μl blood. Unfortunately, it is still not completely known how an untreated HIV-1 infection leads to progressive loss of $CD4^+$ T cells. Interestingly, the rate at which the $CD4^+$ T-cell count declines may be a heritable trait of the virus [18], in a manner that is independent from the heritability of the SPVL. This may lead to evolution of tolerance [186] and an increase of viremic nonprogressors. Hopefully, these scenarios can be explored with a future version of our model.

5.4 Methods

5.4.1 The individual-level model

The differential equations. At the core of the individual-level model, we have a system of ordinary differential equations (ODEs, see Equations (5.1)). The variables in this system of ODEs represent population sizes of quiescent $CD4^+$ T-cells (Q), target cells (T), infected cells (I), $CD8^+$ T cells, i.e. the effector cells (E), and free virus particles (V). When a target cell is infected, it first enters an eclipse phase (I_1), before it starts to produce virions (phase I_2). A graphical representation of the system is given in Figure 5.7A.

The collection of all CTL clonotypes and virus strains form a bipartite graph $G = (\mathscr{V} \cup \mathscr{R}, \mathscr{E})$, with vertices corresponding to immune responses $r \in \mathscr{R}$ and virus strains $v \in \mathscr{V}$, and we draw an edge $(r, v) \in \mathscr{E}$ whenever the response r recognizes an epitope

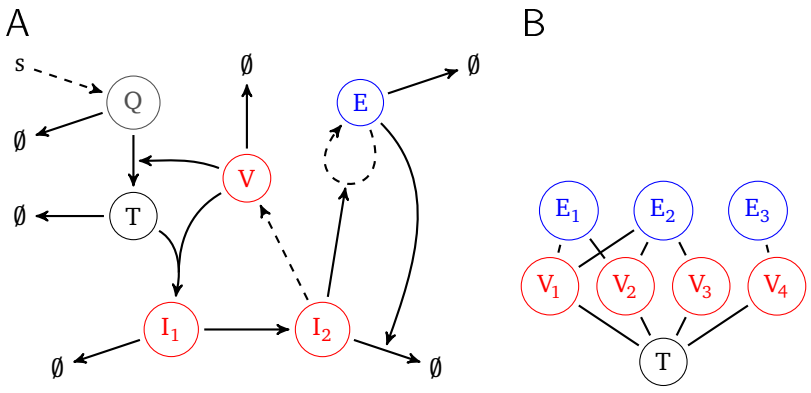


Figure 5.7: Graphical representation of the ODEs. A. In this example, only 1 viral strain and 1 immune response is present. Quiescent $CD4^+$ T cells (Q) are produced from a source (s), and become activated target cells (T) at a rate that is dependent on the number of free virions (V). Target cells can be infected by virions, thereby producing eclipse-phase infected cells (I_1). Eclipse-phase infected cells become virion producing cells (I_2), that enhance the proliferation of effector cells (E). All populations are subject to per-capita death or clearance rates (∅), and the death-rate of virion-producing cells is enhanced by effector cells through killing. Production is indicated by a dashed arrow, while thick arrows signify a transition of catalysis. **B.** Virus strains and immune responses form a bipartite graph. In this example, 4 virus strains are controlled by 3 immune responses. Virus strain 4 is escaped from response 2, but the mutated epitope is recognized by response 3.

of v (see Figure 5.7B). For a vertex x, we write $N(x)$ for the set of neighbors of x.

$$\begin{aligned}
\frac{\mathrm{d}Q}{\mathrm{d}t} &= s-(d_{\mathrm{Q}}+\omega_0)Q-\frac{\sum_{v\in\mathscr{V}}\omega_v V_v}{h_{\mathrm{Q}}+\sum_{v\in\mathscr{V}}V_v}Q\\
\frac{\mathrm{d}T}{\mathrm{d}t} &= \omega_0 Q+\frac{\sum_{v\in\mathscr{V}}\omega_v V_v}{h_{\mathrm{Q}}+\sum_{v\in\mathscr{V}}V_v}Q-d_{\mathrm{T}}T-T\textstyle\sum_{v\in\mathscr{V}}\beta_v V_v\\
\frac{\mathrm{d}I_{1,v}}{\mathrm{d}t} &= f_v\beta_v q_v V_v T-(d_{\mathrm{T}}+\gamma_v)I_{1,v}\\
\frac{\mathrm{d}I_{2,v}}{\mathrm{d}t} &= \gamma_v I_{1,v}-(d_{\mathrm{T}}+\alpha_v)I_{2,v}-m_v I_{2,v}\textstyle\sum_{r\in N(v)}k_r E_r\\
\frac{\mathrm{d}V_v}{\mathrm{d}t} &= p_{\mathrm{V},v}I_{2,v}-(d_{\mathrm{V},v}+\beta_v T)V_v\\
\frac{\mathrm{d}E_r}{\mathrm{d}t} &= p_{\mathrm{E}}\frac{E_r\sum_{v\in N(r)}m_v I_{2,v}}{h_{\mathrm{E},r}+E_r+\sum_{v\in N(r)}I_{2,v}}-d_{\mathrm{E}}E_r\,.
\end{aligned} \tag{5.1}$$

Because the dynamics of free virus particles is fast compared to the dynamics of infected cells, we make the quasi-steady-state (QSS) assumption $\frac{\mathrm{d}V_v}{\mathrm{d}t}=0$, so that

$$\bar{V}_v=\frac{p_{\mathrm{V},v}}{d_{\mathrm{V},v}+\beta_v T}I_{2,v}\,. \tag{5.2}$$

Evolution of the interaction graph. In addition to active responses and the number of infected cells, we also keep track of latent immune responses and all the single-locus mutants of the virus strains present in the system. We think of these as stochastic elements of the system, as they are not described by deterministic ODEs.

Addition of novel viruses and immune responses to the interaction graph. In order to determine when a latent response or a mutant virus should be added to the system, we follow a scheme based on the Sellke algorithm [202]. This scheme consists of the following steps.

We first sample a threshold $\tau\sim\mathrm{Exp}(1)$. We then add a differential equation for a load L to the system, and make the stochastic element deterministic when $L\geq\tau$. The ODE for L equals

$$\frac{\mathrm{d}L}{\mathrm{d}t}=\max\{0,1-1/\mathscr{R}\}\cdot i\,, \tag{5.3}$$

where $\mathscr{R}$ is the reproduction number, and i is the incidence of either cells infected with the mutant, or activated CD8$^+$ T cells. The factor $\max\{0,1-1/\mathscr{R}\}$ in Equation (5.3) equals the so-called "major outbreak" probability of the stochastic population; it is the probability that a birth-death process does not reach the state of 0 individuals. Hence,

symbol	description	h	v	r	value	unit	note
s	production of quiescent target cells	x	.	.	$1.08 \cdot 10^5$	cells d^{-1}	(1)
d_Q	death rate of uninfected quiescent target cells	x	.	.	10^{-3}	d^{-1}	(2)
d_T	death rate of uninfected activated target cells	x	.	.	0.1	d^{-1}	(1)
ω_0	activation rate of quiescent cells	x	.	.	$1.2 \cdot 10^{-2}$	d^{-1}	(2)
ω_v	virus-induced activation rate of quiescent cells	x	x	.	$1.2 \cdot 10^{-8}$	virions$^{-1}d^{-1}$	(2)
h_Q	Michaelis-Menten parameter of quiescent cells	x	.	.	$3.3 \cdot 10^4$	virions	(3)
β_v	infection rate of target cells	x	x	.	$1.37 \cdot 10^{-6}$	virions$^{-1}d^{-1}$	(4)
f_v	fraction of non-abortive target cell infections	x	x	.	0.1	-	(5)
μ_v	mutation probability of a single locus	x	x	.	10^{-6}	-	-
γ_v	eclipse-to-production rate	x	x	.	1.0	d^{-1}	(6)
α_v	virus-induced cytopathic effect	x	x	.	0.9	d^{-1}	(7)
m_v	virus-regulated MHC expression	.	x	.	1.0	-	-
k_r	killing rate of virus-producing cells	.	.	x	$10^{-5} - 10^{-4}$	cells$^{-1}d^{-1}$	-
$p_{V,v}$	production rate of virions	x	x	.	1150	virions d^{-1}	(8)
$d_{V,v}$	clearance rate of virions	x	x	.	23.0	d^{-1}	(8)
p_E	maximal production rate of effector cells	x	.	.	1.1	d^{-1}	(9)
$h_{E,r}$	Michaelis-Menten parameter of effector cells	.	.	x	$10^1 - 10^5$	cells	-
d_E	death rate of effector cells	x	.	.	0.1	d^{-1}	(9)
λ_{naive}	activation rate of naive $CD8^+$ T cells	x	.	.	0.02	d^{-1}	(10)
λ_{memo}	activation rate of memory $CD8^+$ T cells	x	.	.	1.0	d^{-1}	(10)
compound parameters							
q_v	probability of faithful reverse transcription	x	x	.	$e^{-N^*\mu}$	-	(11)
c_v	fraction of infected cells that survive the eclipse phase	x	x	.	$\frac{\gamma_v}{d_T+\gamma_v}$	-	-

Table 5.1: Parameters of the individual-level model. *(continued on page 135)*

Table 5.1: *(continued from page 134)* Parameters are determined by the host (h), the virus (v) or the immune response (r). This is indicated with an "x" in the respective columns. The given parameter value is either the default value, or a range of possible values. Notes: (1) Given $d_T = 0.1$, we choose s such that $T = 10^6$ in the average healthy individual. (2) Taken from Bartha *et al.* [9]. (3) An intermediate virus load to maximize the role of target-cell activation. (4) The parameter β is chosen such that the vRC is 1.5 per day [70]. (5) Conservatively taken from Doitsh *et al.* [50]. (6) Taken from Dixit *et al.* [48]. (7) Chosen such that the death rate of an infected cell is $1d^{-1}$ [142]. (8) The viral clearance rate is taken from Ramratnam *et al.* [180], then we choose a production rate to obtain intermediate virus loads. (9) Given a death rate of $0.1d^{-1}$, the maximum production rate p_E is taken such that the growth rate of E is $1d^{-1}$ [41]. (10) The parameter λ_{naive} is chosen such that viable responses are activated subsequently during the acute phase of the infection, and λ_{memo} is chosen such that viable memory is activated within a few days. (11) By taking $N^* = \frac{1}{2} \cdot \mu^{-1}$, the point-mutation rate ($\mu_v q_v$) is maximized at 2μ.

we correct for the possibility that the newly generated cell clone can go extinct before its population size is large enough to be described by an ODE. Notice that this method is still somewhat phenemenological, as the true emergence dynamics of, e.g., escape mutants likely is much more complex [84].

The incidence of a mutant virus v' is determined by the incidence of its parents, i.e. all virus strains that differ by only one locus from the mutant.

$$i_{v'} = \sum_{v \in \mathcal{V}: d(v,v')=1} \mu_v q_v f_v \beta_v \bar{V}_v T \,, \tag{5.4}$$

where d denotes the Hamming distance between viruses. The reproduction number of the mutant v' equals

$$\mathcal{R}_{v'} = \frac{q_{v'} f_{v'} c_{v'} \beta_{v'} T}{d_T + \alpha_{v'} + m_{v'} \sum_{r \in N_{G'}(v')} k_r E_r} \cdot \frac{p_{V,v'}}{d_{V,v'} + \beta_{v'} T} \,. \tag{5.5}$$

Here, $N_{G'}(v')$ denotes the set of responses in G that recognize an epitope of v'. Notice that the second factor in Equation (5.5) equals the number of virus particles per virion-producing cell. When the virus population becomes more diverse, we have to keep track of a large number of stochastic strains. For instance, the population gets extremely diverse during the acute phase, when the target-cell population is not yet depleted. To keep the simulations computationally feasible, we only follow stochastic strains that are mutants of deterministic viruses whose population sizes have at some point crossed a fixed threshold τ_{mut}. In the simulations, we take $\tau_{mut} = 50$ cells.

A similar procedure is employed to determine when to add a naive or memory response to the system. The incidence of such a latent immune response r' is constant, and

differs only for naive and memory responses:

$$i_{r'} = \begin{cases} \lambda_{\text{naive}} & r' \text{ is a naive response} \\ \lambda_{\text{memo}} & r' \text{ is memory} \end{cases}, \tag{5.6}$$

where we choose $\lambda_{\text{memo}} \gg \lambda_{\text{naive}}$ to express the fact that memory responses are activated more rapidly than naive responses. The reproduction number of a latent response r' equals

$$\mathscr{R}_{r'} = \frac{p_{\text{E}}}{d_{\text{E}}} \frac{\sum_{v \in N_{G'}(r')} m_v I_{2,v}}{h_{r'} + \sum_{v \in N_{G'}(r')} I_{2,v}}. \tag{5.7}$$

Here, $N_{G'}(r')$ denotes the set of viruses in G that have an epitope that r' recognizes.

Deletion of extinct viruses and immune responses from the interaction graph. The ODEs give a deterministic description of the system, that is only a good approximation when the populations are large. When the size of a clone is small, we have to consider the risk of extinction. Whenever a clone size N (i.e. virus I_v or response E_r) is below a threshold τ_{end}, and its reproduction number $\mathscr{R}$ is smaller than 1, the clone is considered endangered. When we find an endangered clone of size $N_0 \approx \tau_{\text{end}}$, we make a prediction about how long it takes for the clone to become extinct—*caeteris paribus*. Given that the clone has per-capita birth rate b and death rate d, a stochastic description of $N \in \mathbb{Z}_{\geq 0}$ is given by the following continuous-time Markov chain

$$N-1 \underset{b\cdot(N-1)}{\overset{d\cdot N}{\rightleftharpoons}} N \underset{b\cdot N}{\overset{d\cdot(N+1)}{\rightleftharpoons}} N+1. \tag{5.8}$$

The goal is to compute the distribution of the "hitting time" θ of the state 0. This can be done by considering the probability generating function (PGF) of N

$$F(t,x) := \sum_{n=0}^{\infty} \mathbb{P}(N(t) = n | N(0) = N_0) x^n, \tag{5.9}$$

which satisfies the following PDE

$$\frac{\partial F}{\partial t} = (x-1)(bx-d)\frac{\partial F}{\partial x}, \tag{5.10}$$

with initial value $F(0,x) = x^{N_0}$. Using the method of characteristics, one finds the following solution

$$F(t,x) = \left[1 - \frac{1-1/\mathscr{R}}{1-\left(1-\frac{1-1/\mathscr{R}}{1-x}\right)\cdot\exp(-(b-d)t)}\right]^{N_0}. \tag{5.11}$$

The distribution of the hitting time θ is now given by

$$\mathbb{P}(\theta \leq t) = \mathbb{P}(N(t) = 0 | N(0) = N_0) = F(t,0) = \left[1 - \frac{1-\mathscr{R}}{\exp(-(b-d)t) - \mathscr{R}}\right]^{N_0} \quad (5.12)$$

Notice that the deterministic description of N is approximately given by $\bar{N}(t) = N_0 \exp((b-d)t)$ and hence we can write

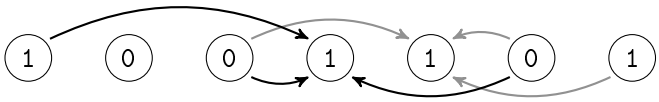
$$\mathbb{P}(\theta \leq t) = \left[1 - \frac{1-\mathscr{R}}{N_0/\bar{N}(t) - \mathscr{R}}\right]^{N_0}. \quad (5.13)$$

In order to decide when to remove a clone from the system of ODEs, we observe that for a random time $t \sim \theta$, the statistic $\mathbb{P}(\theta \leq t|t)$ is uniformly distributed. Hence, we first sample a random deviate $u \sim \text{Uniform}(0,1)$, and then solve Equation (5.13) for $\bar{N}(t)$. We remove the clone whenever its deterministic size drops below the extinction threshold:

$$\bar{N}(t) \leq N_0 \frac{1-u^{1/N_0}}{1-\mathscr{R}u^{1/N_0}}. \quad (5.14)$$

When a clone is endangered, we determine the extinction threshold only once. However, when the environment changes afterwards, the corresponding hitting time would no longer be valid. Since the threshold does not consider a particular time, but involves a population size, this issue is partially resolved. Consider, for example, the situation where a virus strain is endangered because of target-cell limitation. Before the clone reaches the extinction threshold, the target cell population might have recovered, such that the actual growth rate of the endangered strain is larger than the growth rate at the time when the extinction threshold was computed. This means that extinction will occur later than anticipated. This, however, is qualitatively consistent with the stochastic description. Whenever an endangered clone grows above the threshold τ_{end}, we consider it rescued, and if thereafter it falls below the threshold τ_{end} again, we compute a new extinction threshold. In the simulations, we take $\tau_{\text{end}} = 10$ cells.

Host and virus contributions to parameter variation. The parameters of the ODEs depend on virus, immune responses, and the remaining genetic make-up of the host. A parameter x_v that is determined by both host and virus is computed as follows. For a host we draw x_{host} from a Log-Normal distribution with mean x_{default} and a standard deviation of 10% [161]. The virus at node v has a gene that determines a fitness parameter w_x via the NK map (see below). For the virus we compute a parameter $x_{\text{virus}} = x_{\text{default}} \cdot w_x$, whenever a higher value of the parameter corresponds with a higher fitness (e.g. $x = \beta$ or p_V), and $x_{\text{virus}} = x_{\text{default}}/w_x$ when a higher value of the parameter results in a lower viral fitness (e.g. $x = \alpha$ or d_V). The parameter x_v that is used in the system of ODEs is the geometric mean of x_{virus} and x_{host}, i.e., $x_v := \sqrt{x_{\text{virus}} \cdot x_{\text{host}}}$.



b_4	b_1	b_3	b_6	H_4
0	0	0	0	−0.27
⋮	⋮	⋮	⋮	⋮
1	1	0	0	**0.34**
1	1	0	1	0.15

Figure 5.8: Computing a fitness parameter w from a bit string. In this example, we have $N = 7$ and $K = 4$. The black and gray arrows indicate the neighbors of locus 4 and 5, respectively. Locus 4 has neighbors 1, 3 and 6 (i.e. $U_4 = \{4, 1, 3, 6\}$). Given the alleles at locus 4 and its neighbors, we can find that the free energy H_4 for this locus equals 0.34, from a pre-determined table. The total free energy is then given by $H = \frac{1}{N}(H_1 + \cdots + H_7)$, and the fitness parameter equals $w = e^{-(H-H_0)/\kappa}$.

We use the geometric mean instead of the arithmetic mean because otherwise the value of the parameter would be limited from below by the host. For example, when a non-viable mutation would result in $\beta_{\text{virus}} \approx 0$, then we require $\beta_\nu \approx 0$, which is true for the geometric, but not the arithmetic mean.

The parameters of the ODEs that are co-determined by the virus are derived using the NK model that provides a genotype-to-phenotype map [105]. In this model, a genome is modeled using a string of N bits b_ℓ. For each locus ℓ of the genome, a local neighborhood U_ℓ of K loci is chosen, and the alleles b_k at the loci $k \in U_\ell$ together determine the energy $H_\ell = H_\ell((b_k)_{k\in U_\ell})$ associated with locus ℓ. The total energy H of the genome $(b_\ell)_{\ell=1}^N$ is defined as the average of the local energies: $H = \frac{1}{N}\sum_{\ell=1}^N H_\ell$. For each local neighborhood U_ℓ we have to determine 2^K energies, one for each genetic background $(b_k)_{k\in U_\ell} \in \{0,1\}^K$. These values are generated at random at the initialization step of the simulation, and are distributed as $X - \frac{1}{2}$, where $X \sim \text{Beta}(\frac{1}{2}, \frac{1}{2})$.

The neighborhoods U_ℓ are chosen as follows. First, we choose

$$K_\ell \sim \text{Uniform}(1, 2, \ldots, K_{\max}), \tag{5.15}$$

and we always let $\ell \in U_\ell$. Then we choose $K_\ell - 1$ unique neighbors $k = \ell + X$ where $X \sim \text{Skellam}(L, L)$, and $L = 3K_\ell$. We have a separate gene for each parameter, which is a consecutive sub-string of the genome. All genes have the same length N, and only loci in the gene contribute to the parameter-specific energy. The neighborhood of a gene's locus can contain loci of nearby genes on the genome, which facilitates epistatic interactions between genes. Such interactions have for instance been observed between Gag and Protease [63]. The energy H is transformed into to a positive, dimensionless number $w = \exp(-(H - H_0)/\kappa)$, where κ plays the role of the temperature and can be

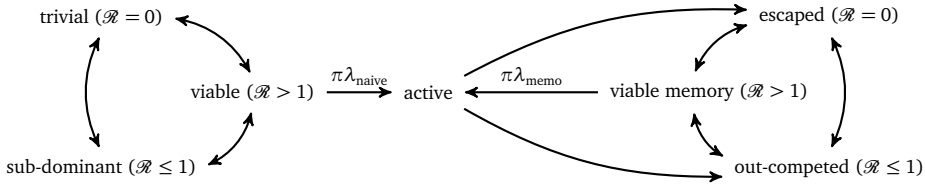


Figure 5.9: The states and transitions of immune responses. Naive responses can toggle between a trivial, sub-dominant and viable state, depending on their reproduction numbers $\mathscr{R}$. An immune response is in the trivial state when none of the strains contain its epitope. Viable responses are activated at rate $\pi\lambda_{\text{naive}}$, where $\pi = 1-1/\mathscr{R}$. Active responses can either be escaped by a fixating escape mutant, or out-competed by a stronger response. In any case the response is now in a memory state. Out-competed responses can be escaped by chance, and memory response can become viable memory when the virus reverts the epitope, or when another strong responses is escaped.

used to tune the ruggedness of the fitness landscape. By trial and error, we found that $\kappa = 0.1$ results in reasonable fitness costs.

The parameter H_0 is used to determine for which energies H the virus-effect on the parameter is neutral (i.e. $w \approx 1$). For large N, we have $\mathbb{E}[H] = 0$, for random genomes, but for evolved genomes, H can be substantially larger than 0. We therefore start with $H_0 = 0$, and sample a large amount of random genomes. These genomes are then evolved using a hill-climbing algorithm, resulting in a sample of evolved energies, and H_0 is henceforth defined as the median of these energies.

Immune responses, epitopes, and HLA alleles. The same string of bits that is used as a model of a genome, is used to encode the epitopes, i.e. the targets of the immune responses. Each immune response r targets an epitope at a fixed locus ℓ_r in the genome. The epitope consists of 3 consecutive bits, where the first bit serves as an anchor position, i.e. this bit models the amino acids that are largely determining binding with the HLA molecule. The remaining bits determine TCR binding. Hence, each epitope has $2^2 = 4$ variants, and at each locus, there are potentially 2 epitopes, one for each anchor bit. In reality, the number of epitopes is much larger, but TCRs are also known to cross-react, i.e. are able to induce a T-cell response to multiple variants of the same epitope, or even multiple epitopes. Additionally, the motifs corresponding to HLA alleles can allow for multiple amino-acids at the anchor positions, which also reduces the effective variability of epitopes.

Immune responses can be shared between individuals, and determine 2 parameters of the system of ODEs. The parameter k_r determines the killing rate of immune response r, and the Michaelis-Menten parameter $h_{\text{E},r}$ determines the amount of antigen that is

required for response r to expand. Since we expect k_r and $h_{\mathrm{E},r}$ to be correlated, we first choose an immuno-dominance parameter $D \sim \mathrm{Beta}(\frac{1}{4}, \frac{3}{4})$, such that sub-dominant responses are more common, and then choose $k_r = k_{\mathrm{sub}}^D \cdot k_{\mathrm{dom}}^{1-D}$ and, likewise $h_{\mathrm{E},r} = h_{\mathrm{sub}}^D \cdot h_{\mathrm{dom}}^{1-D}$, after which we add 10% Log-Normal noise.

A HLA allele is modeled as a set of responses against a fixed set of epitopes. On average, we take 10 epitopes per HLA allele (with a 25% standard deviation), and hence $40 = 2^2 \times 10$ responses against all variants of these epitopes. Each host has $6 = 3 \times 2$ HLA alleles (3 HLA class I loci on each of the 2 chromosomes). Each HLA locus has 30 alleles with a fixed allele frequency distribution, which is sampled from the Dirichlet$(1, \ldots, 1)$ distribution, giving rise to both common and rare alleles.

5.4.2 The population-level model

A dynamic contact network. We simulate a human population where HIV-1 spreads through heterosexual transmission. In our design of the contact network, we aimed for the simplest model that at least has some features that are inherent to real human populations: (i) Individuals have sexual contacts with a limited number of individuals within a time window, but this number differs between individuals, resulting in different risk groups in the population. (ii) Contact formation is assortative as depends typically on age difference, and similarity of promiscuous behavior. (iii) Contact networks display a certain amount of clustering.

The simulated population consists of individuals, or agents, and contacts between these agents (see Figure S5.1). The duration of a contact is finite and Log-Normal-distributed, with a mean and standard deviation of 1y. A subset of the contacts can be used as a transmission route (sexual contacts). Agents have an intrinsic contact-formation rate λ_{cf} (which is Log-Normal-distributed, mean $= 2\mathrm{y}^{-1}$, sd $= 1\mathrm{y}^{-1}$), and attempt to form new contacts according to a Poisson process with rate λ_{cf}. The maximum number of sexual contacts K_{tmr}, differs between individuals and determines their promiscuity. We let $K_{\mathrm{tmr}} \sim$ Pos-Poisson(λ_{cf}), where Pos-Poisson is a zero-truncated Poisson distribution, such that individuals with a higher contact formation rate also tend to have more concurrent sexual contacts.

When an agent A requires a new contact, it first searches for a new partner in their local neighborhood, which is defined as the set of agents B such that A and B share a partner. The probability of contact formation depends on age difference Δa, and the difference in the level of promiscuity ΔK_{tmr}:

$$\mathbb{P}(\text{contact}) = 2^{-(\Delta a/\chi_{\mathrm{age}})^2 - (\Delta K_{\mathrm{tmr}}/\chi_{\mathrm{tmr}})^2}, \tag{5.16}$$

where χ_{age} and χ_{tmr} are the characteristic age difference and promiscuity difference at which contact the contact formation probability $\mathbb{P}(\text{contact})$ is halved. When no

compatible contact was found in the local neighborhood, a new partner is sought in the total population, again giving precedence to contacts with a higher $\mathbb{P}$(contact).

Each agent always tries to maximize its number of sexual contacts. In order to accomplish this, we walk through the population in a random order. For each agent, we test for each contact (again in a random order) if it can become a sexual contact. This is allowed if: (i) the partners have opposite gender, (ii) both partners are older than the minimum age, (iii) both partners have fewer sexual contacts that their limit K_{tmr}.

Demographic turn-over. Susceptible individuals enter the population at a constant rate λ_{birth}. The lifespans of susceptible individuals that are entering the population, are sampled from a realistic parametric distribution with hazard $x \mapsto e^{0.1x-10.5} + e^{-0.4x-8}$ [31]. The initial population is seeded with ages sampled from the distribution that corresponds to the above hazard, assuming a demographic equilibrium. The hazard is also used to compute the birth rate λ_{birth}, such that a virgin population approximately has the desired size.

Transmission and HIV-1 induced mortality. Transmission of the virus is dependent on the virus load (V). The transmission rate β of an infected individual with virus load V equals

$$\beta(V) = \beta_{\text{max}} \frac{V^{\beta_k}}{V^{\beta_k} + \beta_{50}^{\beta_k}} \tag{5.17}$$

where β_{max} is the maximal transmission rate, the parameter β_{50} is the virus load for which the transmission rate is $\frac{1}{2}\beta_{\text{max}}$, and β_k is a shape parameter. The values for these parameters were estimated using a cohort of transmission couples [65, and see Table 5.2]. For determining the moment that a susceptible individual gets infected, we again use the Sellke construction [202]. A susceptible individual i accumulates the sum of all transmission rates from all its sexual contacts j, until the cumulative transmission rate exceeds a pre-determined threshold $\tau_{\text{infect}} \sim \text{Exp}(1)$. More precisely, we define a infection load L via

$$L(t) = \int_0^t \sum_{j \in N_{\text{tmr}}(i,t')} \beta(V_j(t'))\,\mathrm{d}t', \tag{5.18}$$

and infect the susceptible host at time $t_{\text{infect}} := \inf\{t : L(t) \geq \tau_{\text{infect}}\}$. Here, $N_{\text{tmr}}(i,t)$ is the set of all sexual contacts of i at time t. This process is approximated using a τ-leaping algorithm with a time step of 1 week. The transmitter is chosen from all infectious partners with probability proportional to the partners' transmission rates $\beta(V_j)$. From the transmitter, in turn, a virus v is sampled with probability proportional to the clone-specific virus load V_v.

After an acute and asymptomatic infection, an infected agent progresses to AIDS and dies. The duration of the asymptomatic infection depends on the virus load. For a constant virus load V, the average duration of the chronic phase equals

$$D(V) = D_{\max} \frac{V^{D_k} + D_{50}^{D_k}}{D_{50}^{D_k}}, \tag{5.19}$$

Conditioned on the constant virus load, observed lengths of chronic phases show significant variation and can be modeled with a Gamma distribution with mean $D(V)$ and shape parameter ρ. These parameters were estimated from a cohort of untreated individuals [65, and see Table 5.2]. Since the virus load on an infected individual is not constant, we interpret the reciprocal duration $D(V(t))^{-1}$ as a hazard of progressing to AIDS, and remove an infected individual from the population whenever its cumulative hazard $\int_0^t D(V(t'))^{-1} \mathrm{d}t'$ exceeds a pre-determined threshold $\tau_{\text{death}} \sim \text{Gamma}(\rho, \rho^{-1})$.

Notice that an HIV-1 induced death event can be censored by the natural death event of the infected individual, and vice versa. However, all individual-level statistics (such as the SPVL) are always calculated taking the HIV-1 induced death event as the end of the simulation.

To compare our results with the mutation-at-transmission model [206], a virus is given a genome-independent trait g that modulates the virus load. In order to determine AIDS hazard and infectiousness, the virus load $\log_{10}(V)$ is replaced by $\log_{10}(V) + g$. The virus load used for the individual-level model remains unchanged. The trait g is allowed to mutate when the virus is transmitted using mutational variance $\sigma_M = 0.12$ [206], i.e. the transmitted virus samples a $g' \sim \mathcal{N}(g, \sigma_M^2)$.

5.4.3 Associations between HLA alleles and viral polymorphisms

In order to discover associations between HLA alleles and viral polymorphisms, we take a similar approach as, e.g., Carlson *et al.* [29]. Let b_ℓ be the allele at locus ℓ, and let $b_{\ell,\text{WT}}$ be the wild-type allele. We want to test if a mutation at locus ℓ happens preferentially in the context of an HLA allele i. Let $X_i \in \{0, 1\}$ represent the absence or presence of allele i. The probability P_ℓ that $b_\ell \neq b_{\ell,\text{WT}}$ is modeled as

$$\log(P_\ell/(1-P_\ell)) = aX_i + c, \tag{5.20}$$

and parameters a and c are estimated by maximizing the likelihood. The parameter a equals the log-odds that the mutation occurs due to the presence of HLA allele i. The mutation can also be observed because it is present in the transmitted virus. In order to correct for this bias, we define a variable $T_\ell \in \{-1, 1\}$ that equals 1 if the mutation

symbol	description	value	unit	note
$\beta_{\max}$	maximum transmission rate	0.317	y^{-1}	(1)
β_{50}	VL at half maximum transmission rate	13 938	copies/ml	(1)
β_k	shape parameter of the transmission rate	1.02	-	(1)
$D_{\max}$	maximum expected infection length	25.4	y	(1)
D_{50}	VL at half maximum expected infection length	3058	copies/ml	(1)
D_k	shape parameter of the expected infection length	0.41	-	(1)
ρ	shape parameter of the distribution of the infection length	3.46	-	(1)
λ_{cf}	contact formation rate	2	y^{-1}	(2)
χ_{age}	age difference at half maximum contact formation probability	12	y	(2)
χ_{tmr}	promiscuity difference at half maximum contact formation probability	1	partners	(2)
σ_M	mutational variance	0.12	$\log_{10}$ copies/ml	(3)

Table 5.2: Parameters of the population-level model. Notes: The parameters (1) are taken from Fraser *et al.* [65] and (3) Shirreff *et al.* [206]. (2) The sexual contact network are chosen by trial and error, with the requirement that an HIV-1 outbreak is possible.

is transmitted. The corrected model equals

$$\log(P_\ell/(1-P_\ell)) = aX_i + cT_\ell\,. \tag{5.21}$$

For each locus ℓ and HLA allele i, we compute the p-value under the null hypothesis that $a = 0$ using the likelihood-ratio test. Pairs (ℓ, i) for which any of the elements of the contingency table is less than 5 individuals are ignored. In order to correct for multiple testing, the p-values are transformed into q-values using the procedure described by Storey and Tibshirani [215]. An association is called significant if $p < 0.05$ and $q < 0.2$.

5.4.4 Implementation

The model is written in the `C++` programming language and the output of the simulations is analyzed with `Python`. The source code and Jupyter notebooks can be downloaded from `www.github.com/chvandorp/mlev-hiv-model`.

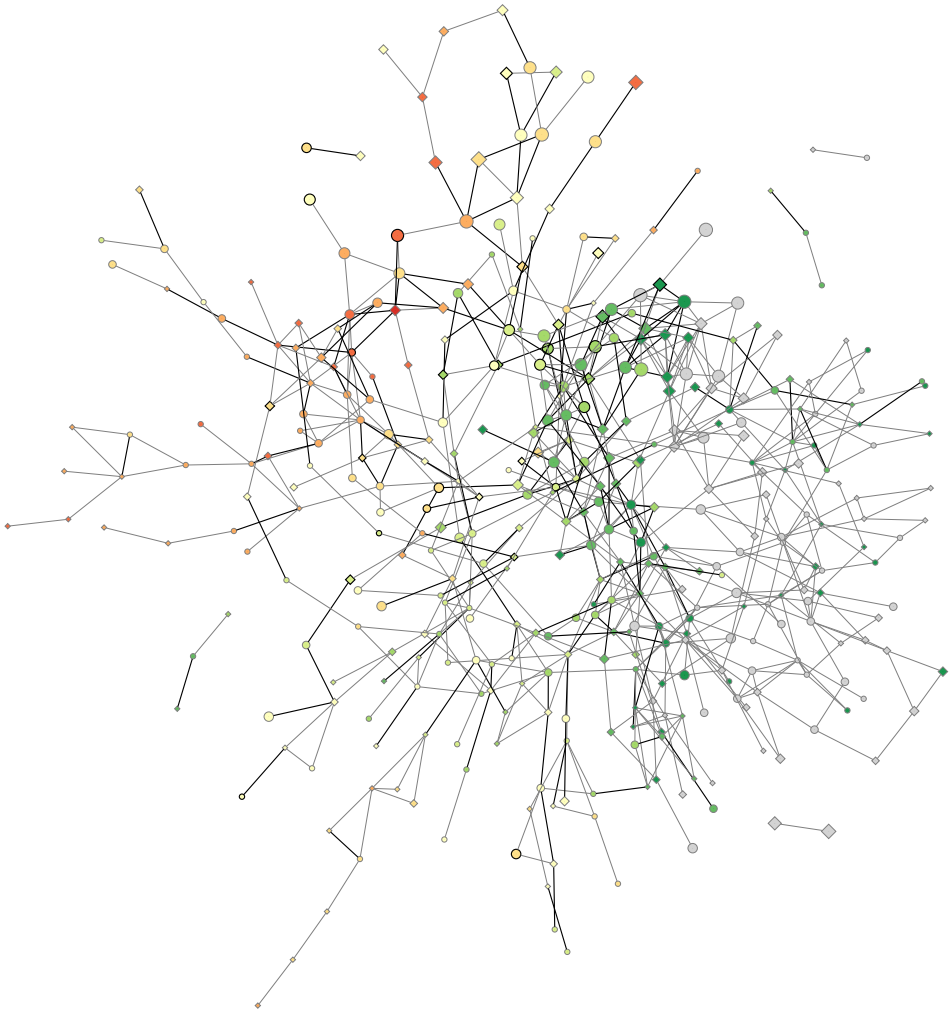


Figure S5.1: Snapshot of the dynamic contact network. Nodes represent the individuals and edges represent contacts between individuals. The heavy edges represent contacts that can be used as transmission routes (sexual contacts), while the light gray edges represent the auxiliary contacts (social contacts). Nodes with a thick outline represent infected agents, and the color of the node corresponds to the age of the individual. The gray nodes are individuals outside the population at risk (individuals of age $\leq$ 15y). The size of the node represents the promiscuity of the individual (the maximum amount of concurrent sexual contacts allowed). The shape of the node represents the gender ($\circ$ for males and $\diamond$ for females).

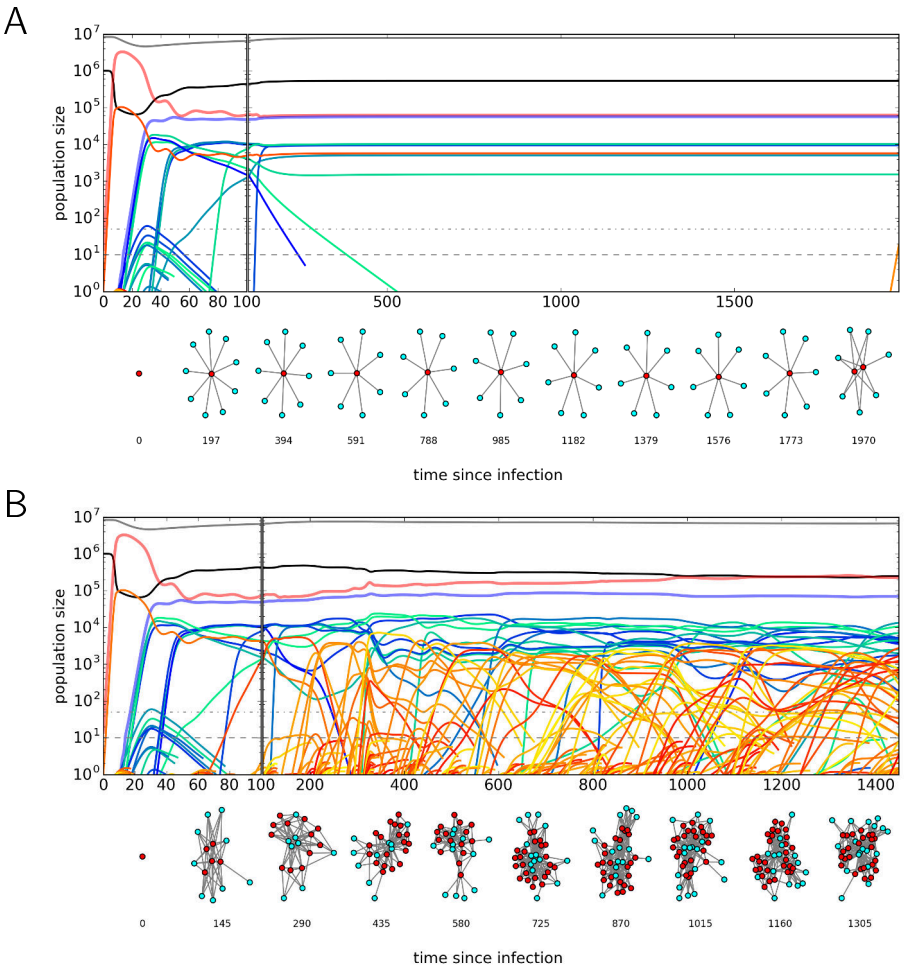


Figure S5.2: Individual-level simulations with low and high mutation rates. The parameters initial conditions are identical those used for Figure 5.1, with the exception of the mutation rate μ, which is tenfold lower in panel **A** ($\mu = 10^{-7}$), and tenfold higher in panel **B** ($\mu = 10^{-5}$).

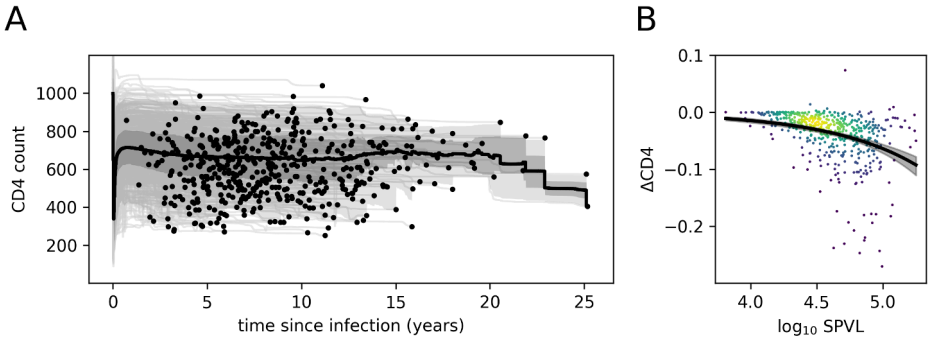


Figure S5.3: Relation between CD4$^+$ T-cell decline and SPVL. The data is taken from a cross-sectional sample of 500 patients, taken 200 years into the epidemic. **A.** A collection of individual-level time series of the CD4$^+$ T-cell count (CD4 count: target cells per µl blood). The gray bands indicate the 2.5, 25, 75, and 97.5 percentiles, the thick line the average. The end of an infection is indicated with a black dot. **B.** The relation between $\log_{10}$ SPVL and the slope of the CD4$^+$ T-cell count (ΔCD4: cells per µl per day). The hue indicates the density of the observations. An exponential curve is fitted through the data (by non-linear least squares). The gray bands indicate the 95% CI (calculated by bootstrapping the data).

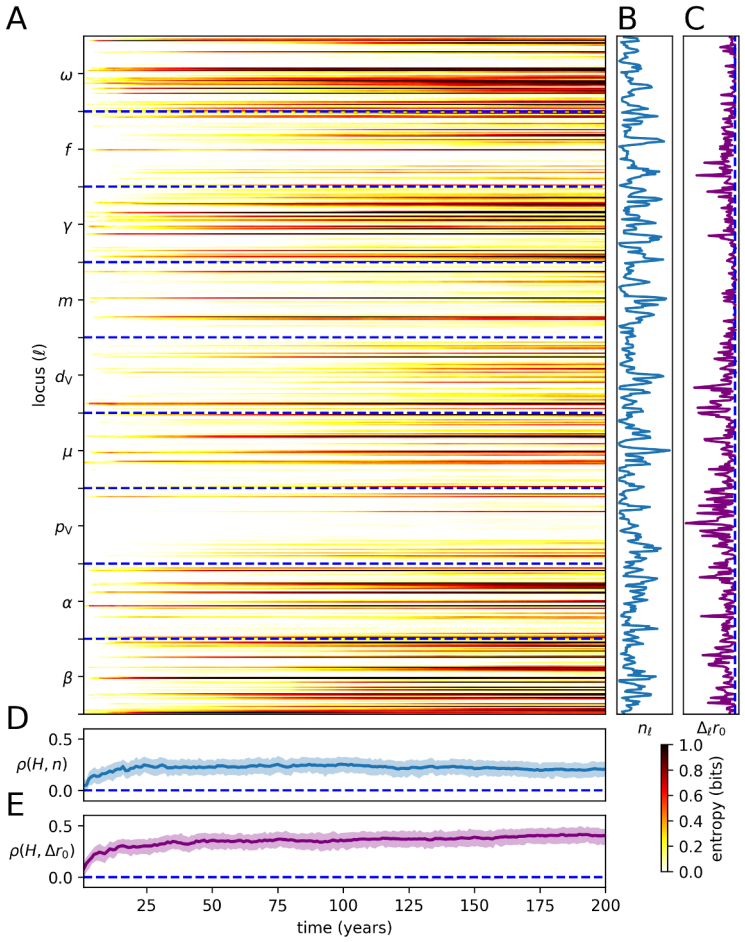


Figure S5.4: Genetic diversity of the virus per viral locus correlates with epitope density and fitness cost. A. For each simulated year and each locus ℓ of the genome, we compute the fraction of genomes b that have allele 1 at locus ℓ, taking all I infected individuals into account: $f_\ell = \frac{1}{I}\sum_b b_\ell$. The corresponding entropy equals $H_\ell = f_\ell \log_2(f_\ell)$. **B.** The number of responses n_ℓ targeting an epitope that contains the locus ℓ is weighted by both the fixed HLA frequency distribution, and the immuno-dominance parameter D. **C.** For each locus ℓ, we consider the replicative capacity of a transmitted virus and a virus with a point mutation at locus ℓ. The relative basic Malthusian fitness ($\Delta_\ell r_0$) is the difference between these vRCs. **D.** The weighted number of responses n_ℓ directed at locus ℓ is correlated with the entropy H_ℓ for each simulated year (using Spearman's ρ). The colored band represents the 95% CI (calculated using bootstrapping). **E.** The relative basic Malthusian fitness (Δr_0) is correlated with the entropy.

Chapter 6

Discussion

In this thesis we have developed statistical and computational models to study evolutionary immuno-epidemiological processes of an acute and a chronic pathogen: IAV and HIV-1. From the perspective of these pathogens, the human host population is highly heterogeneous, which is most notable by the massive polymorphism of the HLA genes. Throughout this thesis, we asked the question if and how a pathogen is able to adapt to such an heterogeneous population.

We have partially answered this question. During the last century, both IAV and HIV-1 have been slowly escaping from CTL responses at a population level. We demonstrated this for IAV in Chapter 2. The effect of recent CTL escapes on the subsequent epidemiological dynamics of IAV appears to be limited. Surprisingly, this is also the case for recent escape from B-cell responses (antibody-antigenic drift), which contradict the current consensus, as the antigenic drift of IAV is thought to be the driver of variation in epidemic size.

Currently, CTL escape also has a limited effect on the virulence of HIV-1, but according to the model developed in Chapter 5, the scale at which immunological pre-adaptation leads to faster disease progression is expected to increase during the next century. Apart from adapting to human immune responses, HIV-1 appears to have evolved (or is currently evolving) its set-point virus load (SPVL) to levels that optimize the population-level fitness ($\mathscr{R}_0$). Our HIV-1 models challenge this idea.

An important piece of evidence for adaptive evolution of the SPVL of HIV-1 is the extensively measured heritability of this SPVL. In Chapter 5 we recreate a scenario in which heritability of a trait as SPVL can arise naturally from ongoing immune escape in a polymorphic host population. At the same time, we underscored in Chapter 4 the possible bias in heritability estimates that can originate from virus-host interactions.

In both cases (IAV and HIV-1), we exposed important gaps in our comprehension of (the impact of) pathogen evolution. Of course, our methods have a number of important

shortcomings, which are ultimately inherent to biological mechanisms that have yet to be discovered.

6.1 Modeling IAV infection histories

In what has been considered a great triumph for the field of applied evolutionary biology, predictive models of antibody-antigenic evolution of the influenza virus are currently used in collaboration with the World Health Organization to inform vaccine updates [157]. From our perspective, an obvious question would be whether it is useful to include T-cell antigenic drift to these predictive models. On the one hand, H3N2 appears to be under positive selection to escape for human CTL responses [140], and loses a CTL epitope every 3 years, as demonstrated in Chapter 2. On the other hand, as we argue in Chapter 3, T-cell antigenic drift seems to have little effect on the susceptibility during subsequent Dutch seasonal influenza epidemics.

This last point should be made with some reservation for three reasons. First, the lack of a relation between antigenic drift and susceptibility is not limited to T-cell antigenic drift, as we are equally unable to explain the variation in susceptibility with antibody-antigenic drift (as calculated by Bedford *et al.* [12]). Apparently, the absence of a direct relation between antigenic drift and variation in susceptibility does not preclude this drift from being useful in a predictive model.

Secondly, influenza-like illness (ILI) is caused not only by the influenza virus, but also by pathogens such as RSV and rhinovirus. The positive predictive value of ILI for an influenza infection is at most 50% [145], which means that many ILI cases in our time series are mis-classified as IAV cases. This mis-qualification might be a significant source of noise, which possibly obscures any signal of antigenic drift. The ILI-classification problem can partially be resolved in the future by incorporating virological data into the Bayesian model. The general practitioners in the sentinel network that contribute the ILI data take two nasal swaps from ILI patients per week, which are sent to the lab for virological testing, making it possible to estimate the weekly fraction of ILI caused by the influenza virus [143], and even the type and subtype [8, 183]. This additional data is only collected in more recent years. Possibly, the type, subtype, sampling date, and the number of sequenced isolates can be used as a substitute for the virological samples for earlier years. We are currently working on the integration of these new data sources in our statistical model.

Thirdly, the currently used measures of antigenic drift might be inadequate for comparison with the complex epidemiological dynamics of influenza. In our Kermack-McKendrick-type model, we stratified the population into 6 age classes with the idea that antigenic drift might be experienced differently by individuals of different ages.

These differences may be the result of prior exposure histories to the virus [36]. For instance, the first HA type (H1, H2 or H3) a person is infected with during childhood, has a strong effect on incidence of, and mortality resulting from avian IAV [81]. This principle can be extended by "charting" the complete infection histories of individuals [64]. One could conceptualize a simple IBM in order to model infection histories (using T-cell epitopes found in historical isolates), and predict susceptibility (distributions) of the population to novel strains, which could be compared to the susceptibility estimated from the ILI time series in the multiple age strata. Moreover, our measure of strain-similarity does not take T-cell cross-reactivity into account [cf. 24], neither are we incorporating uncertainty resulting from unknown immuno-dominance patterns of CTL responses.

If we can resolve the issues outlined above, we might be able to unravel the relation between susceptibility and both antibody- and T-cell antigenic drift. Perhaps there is still a role for T-cell epitopes in influenza vaccine-strain decision making.

6.2 Understanding the pathophysiology of $CD4^+$ T-cell depletion

During untreated HIV infections, the $CD4^+$ T-cell count slowly declines. The rate of this decline correlates with the SPVL, although this correlation is far from perfect [186, 191]. In fact, unexplained variation of the rate of $CD4^+$ T-cell decline has been interpreted as host or virus induced tolerance [18, 186]. AIDS is often defined clinically as a $CD4^+$ T-cell count below 200 cells per μl. Nevertheless, in our HIV-1 models, we have chosen a phenomenological approach in order to model progression to AIDS.

Ideally, we would have used a mechanistic model to describe the progression to AIDS. However, it is often overlooked that the mechanism that causes the $CD4^+$ T-cell decline is still unresolved. At first glance, it seems quite obvious that the fact that HIV-1 infects $CD4^+$ T cells results in a loss of those T cells, but mathematical models have shown that this basic mechanism fails to explain the very slow $CD4^+$ T-cell depletion [244]. In the last decades, a myriad of mathematical models has been developed that can describe slow $CD4^+$ T-cell depletion [2], but none of these models have been accepted as the correct explanation.

Models that can describe a slow $CD4^+$ T-cell decline fall into two categories: evolutionary and ecological models. In the evolutionary models, the $CD4^+$ T-cell decline is somehow linked to viral evolution. The model presented in Chapter 5 falls into this category. The slow decline of the target cell population is due to ongoing escape from immune responses, which results in poorer control of the virus. The decline in $CD4^+$ T cells is therefore closely linked to an increase in virus load. Recently, Bertels *et al.*

[18] have shown that the slope of the CD4$^+$ T-cell count is heritable ($H^2 \approx 17\%$), to a large extent independently from the heritability of the SPVL. This means that our explanation for the target-cell decline cannot be the full mechanism.

Ecological models do not require viral evolution in order to explain the CD4$^+$ T-cell depletion. As the infection of CD4$^+$ cells alone cannot explain the slow depletion of uninfected CD4$^+$ T-cells, some additional biological mechanism is required for these purely ecological models [223]. A mechanism that has gained some popularity recently involves pyroptosis of infected cells. During pyroptosis of abortively infected cells, a cell releases its cytoplasmic content, including inflammatory cytokines. These cytokines can induce pyroptosis in neighboring cells. Wang *et al.* [227] claim that a model with pyroptosis can explain the slow time scale of CD4$^+$ T-cell depletion. What makes ecological models problematic is that often a very slowly changing variable is required to result in a CD4$^+$ T-cell decline that spans years. This slow process is typically phenomenological and experimental evidence is often lacking.

Possibly related to the open questions about disease progression is the still largely unexplained SPVL variation. Low-dimensional ODE models, such as the system used in Chapter 5 have difficulty with explaining this variation [136]. Even by allowing for many parameters of the system of ODEs to vary, we were not able to simulate the true SPVL range of 4–5 orders of magnitude. This means that again a part of the biology is missing from our model.

One interesting idea is that the SPVL and the pace of disease progression is already established during the (hyper) acute phase of the infection. Indeed, Claiborne *et al.* [35] recently demonstrated that the higher vRC of the transmitted virus is related to faster disease progression, independently from the SPVL. Moreover, the vRC of the TF virus is directly related to the size of the initial reservoir, and appears to determine the expression patterns of early inflammatory cytokines. The disruptive impact of the acute infection can in part be prevented by a swift CTL response, as demonstrated by Ndhlovu *et al.* [164], who were able to study VL and CTL dynamics during hyper-acute infections thanks to bi-weekly screening of a high-risk cohort. The idea that the pace of disease progression is determined during the acute phase is further supported by the finding that very early cART can induce post-treatment control in some patients [194], i.e. the VL remained at undetectable levels, even after treatment interruption. Perhaps this early cART has a similar effect as an early CTL response. Such dynamics cannot be generated by the model of Chapter 5. However, incorporating plausible biological features as immune exhaustion can lead to a bi-stable state that describes elite and post-treatment controllers [38].

Nederlandse samenvatting

Pathogenen, zoals virussen en bacteriën, kunnen zich op meerdere niveaus aanpassen aan hun gastheer. Op het tussen-gastheer niveau werkt natuurlijke selectie op infectiviteit, het omzeilen van de immuniteit die aanwezig is in de gastheerpopulatie, en de pathogeen-geïnduceerde mortaliteit. Op het binnen-gastheer niveau worden pathogenen geselecteerd op het ontsnappen aan immuunreacties en de pathogene lading (het aantal pathogenen per eenheid bloed). Al deze vormen van selectiedruk liggen niet noodzakelijkerwijs in elkaars verlengde. Het voorspellen van de uitkomst van selectie op verschillende niveaus wordt nog ingewikkelder gemaakt door de substantiële heterogeniteit binnen de gastheerpopulatie, die hoogstwaarschijnlijk ontstaan is als gevolg van negatieve frequentie-afhankelijke selectie, en de populatie beschermt tegen pathogenen. Als een pathogeen zich namelijk aangepast heeft aan een veelvoorkomend fenotype, heeft een gastheer met een zeldzaam fenotype mogelijk een evolutionair voordeel.

Het meest opvallende voorbeeld van heterogeniteit in de mens is te vinden in de humane leukocytenantigenen (HLA). Deze moleculen zijn te vinden op het oppervlak van de cel en worden gebruikt door het immuunsysteem om peptiden te presenteren aan T lymfocyten. Als de cel geïnfecteerd is met bijvoorbeeld een virus, kan zo'n peptide afkomstig zijn van een viraal proteïne. Een cytotoxische T lymfocyt (CTL) kan dit virale peptide (epitoop) herkennen en de geïnfecteerde cel aanzetten tot apoptose.

In dit proefschrift richten we ons op de immunologie en de epidemiologie van twee pathogenen, één dat acute infecties veroorzaakt (influenza A virus; IAV), en een pathogeen dat leidt tot chronische infecties (humaan immunodeficiëntievirus type 1; HIV-1). In beide gevallen zijn we geïnteresseerd in hoe deze pathogenen evolueren in een heterogene gastheerpopulatie. In het bijzonder bestuderen we hoe deze virussen ontsnappen aan immuunreacties die grotendeels bepaald worden door de zeer polymorfe HLA moleculen, en we bestuderen wat voor gevolgen het ontsnappen aan immuunreacties heeft op het populatieniveau.

De juiste antilichamen tegen IAV kunnen infectie voorkomen doordat ze binden aan het virale oppervlakte-eiwit hemagglutinine (HA), en daarmee de hechting aan de

epitheelcellen in de luchtwegen voorkomen. Doordat er mutaties optreden in HA, kan IAV ontsnappen aan deze antilichamen, en vindt er nagenoeg elke winter een griep epidemie plaats. Deze voortdurende evolutie wordt "antigene drift" genoemd. In het eerste deel van dit proefschrift streven we ernaar om een relatie te vinden tussen antigene drift van seizoensinfluenza en de omvang van de jaarlijkse IAV epidemieën.

In Hoofdstuk 2 bestuderen we de antigene evolutie van IAV. De focus van de meeste voorgaande studies lag op de evolutie van antilichaam-specifieke antigenen (zoals HA). In dit proefschrift besteden we juist meer aandacht aan de CTL epitopen van IAV. We gebruiken deze epitopen, samen met een grote collectie van historische virus sequenties, voor de constructie van een antigene kaart van IAV. Daarnaast laten we zien dat het aantal epitopen in IAV de afgelopen eeuw langzaam maar zeker is afgenomen. Deze afname suggereert dat IAV zich aanpast aan menselijke CTL reacties.

In het derde hoofdstuk kijken we naar het effect van de adaptatie van IAV op populatieniveau. We maken gebruik van data die verzameld is door Nederlandse huisartsen. Door een transmissiemodel zo te kiezen dat het zo goed mogelijk bij de data past, schatten we het aantal vatbare individuen in Nederland aan het begin van 45 influenza epidemieën. Het aantal vatbare individuen varieert sterk tussen de jaren en tussen leeftijdsgroepen. Echter, we vinden geen bewijs voor de hypothese dat deze variatie veroorzaakt wordt door de mate van antigene drift van het virus.

In het tweede deel van dit proefschrift ontwikkelen we twee modellen van de evolutie van HIV-1. Een belangrijk begrip is deze modellen is de SPVL: de virale lading (VL) op het *set-point*. Een HIV-1 infectie begint met een acute fase, waarin de VL snel stijgt. Na een aantal weken is de VL weer gezakt tot een waarde (de SPVL) die jaren (relatief) constant kan blijven. Hoewel de VL binnen een patiënt gelijk blijft, is er tussen patiënten juist een enorme variatie van de SPVL te vinden (5 ordes van grootte). Deze variatie heeft belangrijke gevolgen voor het ziekteverloop en de infectiviteit van (onbehandelde) patiënten. Een patiënt met een lage SPVL is minder infectieus dan iemand met een hoge SPVL. Daartegenover ontwikkelen patiënten met een hoge SPVL sneller AIDS dan patiënten met een lage SPVL. Iemand met een lage SPVL heeft dus meer tijd om het virus over te dragen, maar per tijdseenheid is er minder kans op transmissie. Het product van de tijd tot AIDS en de infectiviteit bepaalt voor een deel hoeveel secundaire infecties veroorzaakt worden door één HIV-1 infectie. Dit product is dus sterk verwant aan de fitness (of het reproductiegetal) van HIV-1 op populatieniveau. Een aantal databronnen suggereert dat de SPVL van HIV-1 geëvolueerd is zo dat deze fitness optimaal is.

Een aanwijzing voor de bovengenoemde hypothese is dat de SPVL overeenkomsten vertoont tussen paren waartussen transmissie van HIV-1 heeft plaatsgevonden. Deze zogenaamde erfelijkheid van de SPVL is een noodzakelijke conditie voor de optimalisatie van de fitness. Eén van onze doelen is om deze optimalisatie te begrijpen in

het kader van een polymorfe gastheerpopulatie. In Hoofdstuk 4 bouwen we een relatief eenvoudig individu-gebaseerd model (IBM) voor de immuno-epidemiologie van HIV-1. Het model maakt gebruik van een fenomenologische beschrijving van de binnen-gastheer evolutie van het virus. Er wordt aangenomen dat de gastheren maximaal heterogeen zijn, hetgeen ons in staat stelt om de representatie van een virus aanzienlijk te vereenvoudigen. Hierdoor kunnen we gebruik maken van de *next-generation matrix* om tijdrovende simulaties te vermijden. Ons model voorspelt niet dat de fitness geoptimaliseerd wordt, maar dat de erfelijkheid van de SPVL voor een groot deel veroorzaakt wordt door een immunologische indruk van het immuunsysteem van de overdragende gastheer.

Aangezien het model dat we ontwikkelen in Hoofdstuk 4 een forse simplificatie van de realiteit is, gebruiken we Hoofdstuk 5 om een veel realistischer HIV-1 model te presenteren, zowel op het tussen- als het binnen-gastheer niveau. Het binnen-gastheer model staat de co-existentie van meerdere HIV-1 stammen toe, alsmede een veelvoud aan immuunreacties. Hiervoor gebruiken we stelsels van gewone differentiaalvergelijkingen. De HIV-1 stammen zijn ieder voorzien van een vereenvoudigd genoom dat voor de parameters in de differentiaalvergelijkingen en een groot aantal mogelijke epitopen codeert. Met dit meer gedetailleerde model bevestigen we een aantal conclusies uit Hoofdstuk 4, en bovendien laat het ons meerdere vormen van adaptatie bestuderen. Recente fylogenetische studies hebben aangetoond dat de overgedragen HIV-1 stam steeds meer is aangepast aan gastheren met bepaalde HLA typen. Deze pre-adaptatie vinden we terug in onze simulaties. Daarnaast kan het model het ontstaan van de erfelijkheid van de SPVL en associaties tussen virale polymorfismen en HLA verklaren.

Acknowledgments

May 15, 2018, two sleepless weeks behind schedule (true to form), I finally sent my manuscript to the assessment committee, consisting of Samuel Alizon, Jonathan Carlson, Mirjam Kretzschmar, Katrina Lythgoe, and Derek Smith. After three anxious weeks, all had agreed to letting me defend the thesis. I am very grateful to the committee members for their thorough reading and motivating recommendation letters.

I sincerely thank my advisers, my *promotor* Rob de Boer and *copromotor* Michiel van Boven. The PhD was not always easy, which could be inferred from the time it has taken me to finish it. However, I look back at many enlightening and motivating meetings, but also dinners, barbecues, conferences, canoeing trips, etc., and I am still astonished by my advisers endless patience. I also have to acknowledge Can Keşmir, for hiring me as a "postdoc", although it took a while to get to the "post" part. Finishing the PhD did not particularly help with advancing our projects, but I hope that we can still successfully finish the ongoing work.

Of course, I am grateful to my paranymphs, Enrico Sandro Colizzi and Juliane Schröter. Sandro has been a supplier of endless (scientific) distraction. Do true RNGs exist? Can violin plots be improved upon? Juliane is continuing the modeling of HIV-1 with Rob. With the aid of Stan (`www.mc-stan.org`) this should be a piece of cake.

I thank Rutger Woolthuis for writing Chapter 2, and his help with Chapter 3, and I hope to be reading his thesis very soon. I am also grateful to Jeffrey Yu for his great work on Chapter 3.

Naturally, I acknowledge my office (Z533) mates: Thomas Cuypers and Folkert de Boer, for putting up with all the stuff about semaphores, and breaking the ice with our great (but short-lived) Duvel-assisted Z533 journal club; Aridaman Pandit, for demonstrating how the word "determine" should be pronounced; and Ewald van Dyk, for being my fellow PhD-less postdoc. Of course I want to thank *all* other members from the Theoretical Biology and Bioinformatics group, but a few of them I want to thank in particuler: Paulien Hogeweg, Jan Kees van Amerongen, Rutger Hermsen, Ronald Bontrop, Renske Vroomans, Hanneke van Deutekom, Jorg Calis, Paola Carrillo-Bustamante, Saikrishna Gadhamsetty, Ioana Niculescu, "pipeline" Bram Gerritsen,

Bram van Dijk, Hilje Doekes, Alessia Peviani, Like Fokkens, and Daniel Weise. I would also like to thank my former colleagues at the RIVM. In particular: Jacco Wallinga, Scott McDonald, Hester Korthals Altes, Annelie Vink, Axel Bonacic Marinovic, and Remko Enserink. Especially, I thank Rolf Ypma, Tjibbe Donker and Dennis te Beest for inviting me as a *junior AIO* on their legendary road trip from Toronto to the Epidemics conference in Boston (and my Canadian aunts for making us breakfast). From the UU (but not TBB or RIVM), I thank KaYin Leung and Sanja Selakovic.

Lastly, I would like to thank my family: My sister Suzanne van Dorp for offering to do the editing of this thesis in MS Word and her last minute help with the *Nederlandse samenvatting*; my brother-in-law Arjan van Laarhoven for his PhD checklists; and my parents Yvonne van Oosterwijk and Harry Rours for not asking too many questions about my progress. Most importantly, I thank my talented wife Doreen Wittenbols for the wonderful book cover, and for all the love and support. Without her this thesis would never have been finished.

Technical and scientific support

A number of people helped me and my co-authors with conducting our research and writing the chapters. First, I thank Jan Kees van Amerongen for providing an excellent computer cluster, and for many instances of helping me with installing all kinds of obscure software.

For help with Chapter 2, I thank Josine van Beek, José Ferreira, Marion Hendriks, Scott McDonald, Adam Meijer, and Johannes Textor. Regarding Chapter 3, I gratefully acknowledge Ingrid Friesema for helping me with the GIS data and Jan van de Kassteele for computing a customized contact matrix. I thank Dennis te Beest and Johannes Textor for technical support, and Jantien Backer for comments on the manuscript. Finally, I am grateful to Mariëtte Hooiveld (NIVEL Primary Care Database, Sentinel Practices, Utrecht, The Netherlands) for her assistance with the ILI data. As to Chapter 4, I thank Paola Carrillo-Bustamante, Rutger Woolthuis and Odo Diekmann for very useful discussions and comments. I gratefully acknowledge Johannes Textor and Hanneke van Deutekom for technical support.

Funding

The research is this thesis was funded by the Complexity(-NET) program of The Netherlands Organisation for Scientific Research (NWO; grant number 645.000.002), NWO (grant number 823.02.014), and the National Institute for Public Health and the Environment (RIVM).

Curriculum vitae

Christiaan H. van Dorp was born in Haarlem in 1986. After attending the Stedelijk Gymnasium Haarlem grammar school, he studied mathematics at the Universiteit van Amsterdam (UvA). In 2011, he graduated *cum laude* at the UvA and received his MSc in mathematics. For his thesis on "Vector-valued Siegel modular forms of genus 2", he was awarded the student prize from the Geometry and Quantum Theory cluster of The Netherlands Organisation for Scientific Research (NWO). From 2011 to 2015, Christiaan worked as a PhD student at the Theoretical Biology and Bioinformatics group at Universiteit Utrecht (UU), and the National Institute for Public Health and the Environment (RIVM), under supervision of Prof.dr. Rob J. de Boer and Dr. Michiel van Boven. From 2016 to 2018, Christiaan worked as a researcher at the UU under supervision of Dr. Can Keşmir on Bayesian models of HLA and KIR associations with various disease traits.

Publications

1. C. H. van Dorp, M. van Boven, and R. J. de Boer. Immuno-epidemiological modeling of HIV-1 predicts high heritability of the set-point virus load, while selection for CTL escape dominates virulence evolution. *PLOS Computational Biology*, 10(12):e1003899, 2014

2. R. G. Woolthuis, C. H. van Dorp, C. Keşmir, R. J. de Boer, and M. van Boven. Long-term adaptation of the influenza A virus by escaping cytotoxic T-cell recognition. *Nature Scientific Reports*, 6:33334, 2016

3. M. van Boven, J. van de Kassteele, M. J. Korndewal, C. H. van Dorp, M. Kretzschmar. F. van der Klis, H. E. de Melker, A. C. Vossen, and D. van Baarle. Infectious reactivation of cytomegalovirus explaining age- and sex-specific patterns of seroprevalence. *PLOS Computational Biology*, 13(9):e1005719, 2017

Bibliography

[1] Alizon, S. and C. Fraser. Within-host and between-host evolutionary rates across the HIV-1 genome. *Retrovirology*, 10:49, 2013.

[2] Alizon, S. and C. Magnus. Modelling the course of an HIV infection: insights from ecology and evolution. *Viruses*, 4(10):1984–2013, 2012.

[3] Alizon, S., V. von Wyl, T. Stadler, R. D. Kouyos, S. Yerly, *et al.* Phylogenetic Approach Reveals That Virus Genotype Largely Determines HIV Set-Point Viral Load. *PLoS Pathog*, 6(9):e1001123, 2010.

[4] Allen, T. M., M. Altfeld, X. G. Yu, K. M. O'Sullivan, M. Lichterfeld, *et al.* Selection, Transmission, and Reversion of an Antigen-Processing Cytotoxic T-Lymphocyte Escape Mutation in Human Immunodeficiency Virus Type 1 Infection. *Journal of Virology*, 78(13):7069–7078, 2004.

[5] Althaus, C. L. and R. J. De Boer. Dynamics of immune escape during HIV/SIV infection. *PLoS Comput. Biol.*, 4(7):e1000103, 2008.

[6] Asquith, B., C. T. T. Edwards, M. Lipsitch, and A. R. McLean. Inefficient Cytotoxic T Lymphocyte–Mediated Killing of HIV-1–Infected Cells In Vivo. *PLoS Biol*, 4(4):e90, 2006.

[7] Assarsson, E., H. H. Bui, J. Sidney, Q. Zhang, J. Glenn, *et al.* Immunomic analysis of the repertoire of T-cell specificities for influenza A virus in humans. *J. Virol.*, 82(24):12241–12251, 2008.

[8] Baguelin, M., S. Flasche, A. Camacho, N. Demiris, E. Miller, *et al.* Assessing optimal target populations for influenza vaccination programmes: an evidence synthesis and modelling study. *PLoS Med.*, 10(10):e1001527, 2013.

[9] Bartha, I., P. Simon, and V. Müller. Has HIV evolved to induce immune pathogenesis? *Trends Immunol.*, 29(7):322–328, 2008.

[10] Barton, J. P., M. Kardar, and A. K. Chakraborty. Scaling laws describe memories of host-pathogen riposte in the HIV population. *Proc. Natl. Acad. Sci. U.S.A.*, 112(7):1965–1970, 2015.

[11] Bedford, T., A. Rambaut, and M. Pascual. Canalization of the evolutionary trajectory of the human influenza virus. *BMC Biol.*, 10:38, 2012.

[12] Bedford, T., M. A. Suchard, P. Lemey, G. Dudas, V. Gregory, *et al.* Integrating influenza antigenic dynamics with molecular evolution. *Elife*, 3:e01914, 2014.

[13] te Beest, D. E., M. van Boven, M. Hooiveld, C. van den Dool, and J. Wallinga. Driving factors of influenza transmission in the Netherlands. *Am. J. Epidemiol.*, 178(9):1469–1477, 2013.

[14] Begon, M., M. Bennett, R. G. Bowers, N. P. French, S. M. Hazel, *et al.* A clarification of transmission terms in host-microparasite models: numbers, densities and areas. *Epidemiol. Infect.*, 129(1):147–153, 2002.

[15] Berkhoff, E. G., A. C. Boon, N. J. Nieuwkoop, R. A. Fouchier, K. Sintnicolaas, *et al.* A mutation in the HLA-B*2705-restricted NP383-391 epitope affects the human influenza A virus-specific cytotoxic T-lymphocyte response in vitro. *J. Virol.*, 78(10):5216–5222, 2004.

[16] Berkhoff, E. G., E. de Wit, M. M. Geelhoed-Mieras, A. C. Boon, J. Symons, *et al.* Fitness costs limit escape from cytotoxic T lymphocytes by influenza A viruses. *Vaccine*, 24(44-46):6594–6596, 2006.

[17] Berlanda Scorza, F., V. Tsvetnitsky, and J. J. Donnelly. Universal influenza vaccines: Shifting to better vaccines. *Vaccine*, 2016.

[18] Bertels, F., A. Marzel, G. Leventhal, V. Mitov, J. Fellay, *et al.* Dissecting HIV Virulence: Heritability of Setpoint Viral Load, CD4+ T-Cell Decline, and Per-Parasite Pathogenicity. *Mol. Biol. Evol.*, 35(1):27–37, 2018.

[19] Blanquart, F., M. K. Grabowski, J. Herbeck, F. Nalugoda, D. Serwadda, *et al.* A transmission-virulence evolutionary trade-off explains attenuation of hiv-1 in uganda. *eLife*, 5:e20492, 2016.

[20] Blanquart, F., C. Wymant, M. Cornelissen, A. Gall, M. Bakker, *et al.* Viral genetic variation accounts for a third of variability in HIV-1 set-point viral load in Europe. *PLoS Biol.*, 15(6):e2001855, 2017.

[21] Bogner, P., I. Capua, D. J. Lipman, and N. J. Cox. A global initiative on sharing avian flu data. *Nature*, 442:981, 2006.

[22] Bonhoeffer, S., C. Fraser, and G. E. Leventhal. High heritability is compatible with the broad distribution of set point viral load in HIV carriers. *PLoS Pathog.*, 11(2):e1004634, 2015.

[23] Bonhoeffer, S., G. A. Funk, H. F. Gunthard, M. Fischer, and V. Muller. Glancing behind virus load variation in HIV-1 infection. *Trends Microbiol.*, 11(11):499–504, 2003.

[24] Boon, A. C., G. de Mutsert, D. van Baarle, D. J. Smith, A. S. Lapedes, *et al.* Recognition of homo- and heterosubtypic variants of influenza A viruses by human CD8+ T lymphocytes. *J. Immunol.*, 172(4):2453–2460, 2004.

[25] Bui, H. H., B. Peters, E. Assarsson, I. Mbawuike, and A. Sette. Ab and T cell epitopes of influenza A virus, knowledge and opportunities. *Proc. Natl. Acad. Sci. U.S.A.*, 104(1):246–251, 2007.

[26] Carlson, J. M., Z. L. Brumme, C. M. Rousseau, C. J. Brumme, P. Matthews, *et al.* Phylogenetic dependency networks: inferring patterns of CTL escape and codon covariation in HIV-1 Gag. *PLoS Comput. Biol.*, 4(11):e1000225, 2008.

[27] Carlson, J. M., V. Y. Du, N. Pfeifer, A. Bansal, V. Y. Tan, *et al.* Impact of pre-adapted HIV transmission. *Nat. Med.*, 22(6):606–613, 2016.

[28] Carlson, J. M., A. Q. Le, A. Shahid, and Z. L. Brumme. HIV-1 adaptation to HLA: a window into virus-host immune interactions. *Trends Microbiol.*, 23(4):212–224, 2015.

[29] Carlson, J. M., J. Listgarten, N. Pfeifer, V. Tan, C. Kadie, *et al.* Widespread impact of HLA restriction on immune control and escape pathways of HIV-1. *J. Virol.*, 86(9):5230–5243, 2012.

[30] Carlson, J. M., M. Schaefer, D. C. Monaco, R. Batorsky, D. T. Claiborne, *et al.* HIV transmission. Selection bias at the heterosexual HIV-1 transmission bottleneck. *Science*, 345(6193):1254031, 2014.

[31] Carnes, B. A., L. R. Holden, S. J. Olshansky, M. T. Witten, and J. S. Siegel. Mortality partitions and their relevance to research on senescence. *Biogerontology*, 7(4):183–198, 2006.

[32] Carrat, F., E. Vergu, N. M. Ferguson, M. Lemaitre, S. Cauchemez, *et al.* Time lines of infection and disease in human influenza: a review of volunteer challenge studies. *Am. J. Epidemiol.*, 167(7):775–785, 2008.

[33] Childs, L. M., E. B. Baskerville, and S. Cobey. Trade-offs in antibody repertoires to complex antigens. *Philos. Trans. R. Soc. Lond., B, Biol. Sci.*, 370(1676), 2015.

[34] Chowell, G., M. A. Miller, and C. Viboud. Seasonal influenza in the United States, France, and Australia: transmission and prospects for control. *Epidemiol. Infect.*, 136(6):852–864, 2008.

[35] Claiborne, D. T., J. L. Prince, E. Scully, G. Macharia, L. Micci, *et al.* Replicative fitness of transmitted HIV-1 drives acute immune activation, proviral load in memory CD4+ T cells, and disease progression. *Proc. Natl. Acad. Sci. U.S.A.*, 112(12):E1480–1489, 2015.

[36] Cobey, S. and S. E. Hensley. Immune history and influenza virus susceptibility. *Curr Opin Virol*, 22:105–111, 2017.

[37] Cobey, S. and M. Pascual. Consequences of host heterogeneity, epitope immunodominance, and immune breadth for strain competition. *J. Theor. Biol.*, 270(1):80–87, 2011.

[38] Conway, J. M. and A. S. Perelson. Post-treatment control of HIV infection. *Proc. Natl. Acad. Sci. U.S.A.*, 112(17):5467–5472, 2015.

[39] Cotton, L. A., X. T. Kuang, A. Q. Le, J. M. Carlson, B. Chan, *et al.* Genotypic and functional impact of HIV-1 adaptation to its host population during the North American epidemic. *PLoS Genet.*, 10(4):e1004295, 2014.

[40] Cummins, N. W. and A. D. Badley. Making sense of how HIV kills infected CD4 T cells: implications for HIV cure. *Mol Cell Ther*, 2:20, 2014.

[41] Davenport, M. P., R. M. Ribeiro, and A. S. Perelson. Kinetics of virus-specific CD8+ T cells and the control of human immunodeficiency virus infection. *J. Virol.*, 78(18):10096–10103, 2004.

[42] Day, T. and S. Gandon. Applying population-genetic models in theoretical evolutionary epidemiology. *Ecol. Lett.*, 10(10):876–888, 2007.

[43] van Deutekom, H. W., I. Hoof, R. E. Bontrop, and C. Keşmir. A comparative analysis of viral peptides presented by contemporary human and chimpanzee MHC class I molecules. *J. Immunol.*, 187(11):5995–6001, 2011.

[44] van Deutekom, H. W., G. Wijnker, and R. J. de Boer. The rate of immune escape vanishes when multiple immune responses control an HIV infection. *J. Immunol.*, 191(6):3277–3286, 2013.

[45] Diekmann, O., H. Heesterbeek, and T. Britton. *Mathematical Tools for Understanding Infectious Disease Dynamics*. Princeton University Press, 2013.

[46] Diekmann, O., J. A. P. Heesterbeek, and J. A. J. Metz. On the definition and the computation of the basic reproduction ratio R 0 in models for infectious diseases in heterogeneous populations. *Journal of Mathematical Biology*, 28(4):365–382, 1990.

[47] Dijkstra, F., G. A. Donker, B. Wilbrink, A. B. Van Gageldonk-Lafeber, and M. A. Van Der Sande. Long time trends in influenza-like illness and associated determinants in The Netherlands. *Epidemiol. Infect.*, 137(4):473–479, 2009.

[48] Dixit, N. M., M. Markowitz, D. D. Ho, and A. S. Perelson. Estimates of intracellular delay and average drug efficacy from viral load data of HIV-infected individuals under antiretroviral therapy. *Antivir. Ther. (Lond.)*, 9(2):237–246, 2004.

[49] Doekes, H. M., C. Fraser, and K. A. Lythgoe. Effect of the Latent Reservoir on the Evolution of HIV at the Within- and Between-Host Levels. *PLoS Comput. Biol.*, 13(1):e1005228, 2017.

[50] Doitsh, G., N. L. Galloway, X. Geng, Z. Yang, K. M. Monroe, *et al.* Cell death by pyroptosis drives CD4 T-cell depletion in HIV-1 infection. *Nature*, 505(7484):509–514, 2014.

[51] Dong, T., Y. Zhang, K. Y. Xu, H. Yan, I. James, *et al.* Extensive HLA-driven viral diversity following a narrow-source HIV-1 outbreak in rural China. *Blood*, 118(1):98–106, 2011.

[52] Donker, G. A. *NIVEL Primary Care Database - Sentinel Practices 2015*. NIVEL, Netherlands Institute for Health Services Research, Utrecht, 2016.

[53] Dorak, M. T., J. Tang, A. Penman-Aguilar, A. O. Westfall, I. Zulu, *et al.* Transmission of HIV-1 and HLA-B allele-sharing within serodiscordant heterosexual Zambian couples. *Lancet*, 363(9427):2137–2139, 2004.

[54] van Dorp, C. H., M. van Boven, and R. J. de Boer. Immuno-epidemiological modeling of HIV-1 predicts high heritability of the set-point virus load, while selection for CTL escape dominates virulence evolution. *PLoS Comput. Biol.*, 10(12):e1003899, 2014.

[55] van den Driessche, P. and J. Watmough. Reproduction numbers and sub-threshold endemic equilibria for compartmental models of disease transmission. *Math Biosci*, 180:29–48, 2002.

[56] Epstein, S. L. Prior H1N1 influenza infection and susceptibility of Cleveland Family Study participants during the H2N2 pandemic of 1957: an experiment of nature. *J. Infect. Dis.*, 193(1):49–53, 2006.

[57] Fellay, J., D. Ge, K. V. Shianna, S. Colombo, B. Ledergerber, *et al.* Common Genetic Variation and the Control of HIV-1 in Humans. *PLoS Genet*, 5(12):e1000791, 2009.

[58] Fellay, J., K. V. Shianna, D. Ge, S. Colombo, B. Ledergerber, *et al.* A Whole-Genome Association Study of Major Determinants for Host Control of HIV-1. *Science*, 317(5840):944–947, 2007.

[59] Ferguson, A. L., J. K. Mann, S. Omarjee, T. Ndung'u, B. D. Walker, *et al.* Translating HIV sequences into quantitative fitness landscapes predicts viral vulnerabilities for rational immunogen design. *Immunity*, 38(3):606–617, 2013.

[60] Fernandez, C. S., M. Z. Smith, C. J. Batten, R. D. Rose, J. C. Reece, *et al.* Vaccine-Induced T Cells Control Reversion of AIDS Virus Immune Escape Mutants. *Journal of Virology*, 81(8):4137–4144, 2007.

[61] Fideli, Ü. S., S. A. Allen, R. Musonda, S. Trask, B. H. Hahn, *et al.* Virologic and immunologic determinants of heterosexual transmission of human immunodeficiency virus type 1 in africa. *AIDS Research and Human Retroviruses*, 17(10):901–910, 2001.

[62] Fleming, D. M. and A. J. Elliot. Lessons from 40 years' surveillance of influenza in England and Wales. *Epidemiol. Infect.*, 136(7):866–875, 2008.

[63] Flynn, W. F., M. W. Chang, Z. Tan, G. Oliveira, J. Yuan, *et al.* Deep sequencing of protease inhibitor resistant HIV patient isolates reveals patterns of correlated mutations in Gag and protease. *PLoS Comput. Biol.*, 11(4):e1004249, 2015.

[64] Fonville, J. M., S. H. Wilks, S. L. James, A. Fox, M. Ventresca, *et al.* Antibody landscapes after influenza virus infection or vaccination. *Science*, 346(6212):996–1000, 2014.

[65] Fraser, C., T. D. Hollingsworth, R. Chapman, F. de Wolf, and W. P. Hanage. Variation in HIV-1 set-point viral load: epidemiological analysis and an evolutionary hypothesis. *Proc. Natl. Acad. Sci. U.S.A.*, 104(44):17441–17446, 2007.

[66] Fraser, C., K. Lythgoe, G. E. Leventhal, G. Shirreff, T. D. Hollingsworth, *et al.* Virulence and pathogenesis of HIV-1 infection: an evolutionary perspective. *Science*, 343(6177):1243727, 2014.

[67] Friesema, I. H., C. E. Koppeschaar, G. A. Donker, F. Dijkstra, S. P. van Noort, *et al.* Internet-based monitoring of influenza-like illness in the general population: experience of five influenza seasons in The Netherlands. *Vaccine*, 27(45):6353–6357, 2009.

[68] Fryer, H. R., J. Frater, A. Duda, D. Palmer, R. E. Phillips, *et al.* Cytotoxic T-lymphocyte escape mutations identified by HLA association favor those which escape and revert rapidly. *J. Virol.*, 86(16):8568–8580, 2012.

[69] Fryer, H. R., J. Frater, A. Duda, M. G. Roberts, R. E. Phillips, *et al.* Modelling the evolution and spread of HIV immune escape mutants. *PLoS Pathog.*, 6(11):e1001196, 2010.

[70] Gadhamsetty, S., T. Coorens, and R. J. de Boer. Notwithstanding Circumstantial Alibis, Cytotoxic T Cells Can Be Major Killers of HIV-1-Infected Cells. *J. Virol.*, 90(16):7066–7083, 2016.

[71] Garcia, V. and M. W. Feldman. Within-Epitope Interactions Can Bias CTL Escape Estimation in Early HIV Infection. *Front Immunol*, 8:423, 2017.

[72] Garcia, V., M. W. Feldman, and R. R. Regoes. Investigating the Consequences of Interference between Multiple CD8+ T Cell Escape Mutations in Early HIV Infection. *PLoS Comput. Biol.*, 12(2):e1004721, 2016.

[73] Gaufin, T., R. M. Ribeiro, R. Gautam, J. Dufour, D. Mandell, *et al.* Experimental depletion of CD8+ cells in acutely SIVagm-Infected African Green Monkeys results in increased viral replication. *Retrovirology*, 7(1):42, 2010.

[74] Gelman, A., J. Hwang, and A. Vehtari. Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, 24(6):997–1016, 2014.

[75] Goepfert, P. A., W. Lumm, P. Farmer, P. Matthews, A. Prendergast, *et al.* Transmission of HIV-1 Gag immune escape mutations is associated with reduced viral load in linked recipients. *The Journal of Experimental Medicine*, 205(5):1009–1017, 2008.

[76] Goeyvaerts, N., L. Willem, K. Van Kerckhove, Y. Vandendijck, G. Hanquet, *et al.* Estimating dynamic transmission model parameters for seasonal influenza by fitting to age and season-specific influenza-like illness incidence. *Epidemics*, 13:1–9, 2015.

[77] Gog, J. R., G. F. Rimmelzwaan, A. D. Osterhaus, and B. T. Grenfell. Population dynamics of rapid fixation in cytotoxic T lymphocyte escape mutants of influenza A. *Proc. Natl. Acad. Sci. U.S.A.*, 100(19):11143–11147, 2003.

[78] Goldfield, M., J. D. Bartley, W. Pizzuti, H. C. Black, R. Altman, *et al.* Influenza in New Jersey in 1976: isolations of influenza A/New Jersey/76 virus at Fort Dix. *J. Infect. Dis.*, 136 Suppl:S347–355, 1977.

[79] Gong, L. I. and J. D. Bloom. Epistatically interacting substitutions are enriched during adaptive protein evolution. *PLoS Genet.*, 10(5):e1004328, 2014.

[80] Goonetilleke, N., M. K. P. Liu, J. F. Salazar-Gonzalez, G. Ferrari, E. Giorgi, *et al.* The first T cell response to transmitted/founder virus contributes to the control of acute viremia in HIV-1 infection. *The Journal of Experimental Medicine*, 206(6):1253–1272, 2009.

[81] Gostic, K. M., M. Ambrose, M. Worobey, and J. O. Lloyd-Smith. Potent protection against H5N1 and H7N9 influenza via childhood hemagglutinin imprinting. *Science*, 354(6313):722–726, 2016.

[82] Goulder, P. J. R. and D. I. Watkins. Impact of MHC class I diversity on immune control of immunodeficiency virus replication. *Nature Reviews Immunology*, 8(8):619–630, 2008.

[83] Han, C., A. Kawana-Tachikawa, A. Shimizu, D. Zhu, H. Nakamura, *et al.* Switching and emergence of CTL epitopes in HIV-1 infection. *Retrovirology*, 11:38, 2014.

[84] Hartfield, M. and S. Alizon. Within-host stochastic emergence dynamics of immune-escape mutants. *PLoS Comput. Biol.*, 11(3):e1004149, 2015.

[85] Hayward, A. C., L. Wang, N. Goonetilleke, E. B. Fragaszy, A. Bermingham, *et al.* Natural T Cell-mediated Protection against Seasonal and Pandemic Influenza. Results of the Flu Watch Cohort Study. *Am. J. Respir. Crit. Care Med.*, 191(12):1422–1431, 2015.

[86] Hecht, F. M., W. Hartogensis, L. Bragg, P. Bacchetti, R. Atchison, *et al.* HIV RNA level in early infection is predicted by viral load in the transmission source. *AIDS April 24, 2010*, 24(7):941–945, 2010.

[87] Henn, M. R., C. L. Boutwell, P. Charlebois, N. J. Lennon, K. A. Power, *et al.* Whole Genome Deep Sequencing of HIV-1 Reveals the Impact of Early Minor Variants Upon Immune Recognition During Acute Infection. *PLoS Pathog*, 8(3):e1002529, 2012.

[88] Herbeck, J. T., J. E. Mittler, G. S. Gottlieb, S. M. Goodreau, J. T. Murphy, *et al.* Evolution of hiv virulence in response to widespread scale up of antiretroviral therapy: a modeling study. *Virus Evolution*, 2(2), 2016.

[89] Herbeck, J. T., J. E. Mittler, G. S. Gottlieb, and J. I. Mullins. An HIV epidemic model based on viral load dynamics: value in assessing empirical trends in HIV virulence and community viral load. *PLoS Comput. Biol.*, 10(6):e1003673, 2014.

[90] Herbeck, J. T., V. Muller, B. S. Maust, B. Ledergerber, C. Torti, *et al.* Is the virulence of HIV changing? A meta-analysis of trends in prognostic markers of HIV disease progression and transmission. *AIDS*, 26(2):193–205, 2012.

[91] Herfst, S., M. Imai, Y. Kawaoka, and R. A. Fouchier. Avian influenza virus transmission to mammals. *Curr. Top. Microbiol. Immunol.*, 385:137–155, 2014.

[92] Hertz, T., C. M. Oshansky, P. L. Roddam, J. P. DeVincenzo, M. A. Caniza, *et al.* HLA targeting efficiency correlates with human T-cell response magnitude and with mortality from influenza A infection. *Proc. Natl. Acad. Sci. U.S.A.*, 110(33):13492–13497, 2013.

[93] Hill, A. L., D. I. Rosenbloom, and M. A. Nowak. Evolutionary dynamics of HIV at multiple spatial and temporal scales. *J. Mol. Med.*, 90(5):543–561, 2012.

[94] Hinkley, T., J. Martins, C. Chappey, M. Haddad, E. Stawiski, *et al.* A systems analysis of mutational effects in HIV-1 protease and reverse transcriptase. *Nat. Genet.*, 43(5):487–489, 2011.

[95] Hodcroft, E., E. Fearnhill, A. Phillips, D. Dunn, D. Pillay, *et al.* Viral Genotype in Subtype C Significantly Influences Plasma Viral Load. 2013. Poster presentation 258b at CROI 2013.

[96] Hodcroft, E., J. D. Hadfield, E. Fearnhill, A. Phillips, D. Dunn, *et al.* The Contribution of Viral Genotype to Plasma Viral Set-Point in HIV Infection. *PLoS Pathog.*, 10(5):e1004112, 2014.

[97] Hollingsworth, T. D., R. M. Anderson, and C. Fraser. HIV-1 transmission, by stage of infection. *J. Infect. Dis.*, 198(5):687–693, 2008.

[98] Hollingsworth, T. D., O. Laeyendecker, G. Shirreff, C. A. Donnelly, D. Serwadda, *et al.* HIV-1 transmitting couples have similar viral load set-points in Rakai, Uganda. *PLoS Pathog.*, 6(5):e1000876, 2010.

[99] Hoof, I., B. Peters, J. Sidney, L. E. Pedersen, A. Sette, *et al.* NetMHCpan, a method for MHC class I binding prediction beyond humans. *Immunogenetics*, 61(1):1–13, 2009.

[100] Hool, A., G. E. Leventhal, and S. Bonhoeffer. Virus-induced target cell activation reconciles set-point viral load heritability and within-host evolution. *Epidemics*, 5(4):174–180, 2013.

[101] Jin, X., D. E. Bauer, S. E. Tuttleton, S. Lewin, A. Gettie, *et al.* Dramatic Rise in Plasma Viremia after CD8+ T Cell Depletion in Simian Immunodeficiency Virus–infected Macaques. *The Journal of Experimental Medicine*, 189(6):991–998, 1999.

[102] Kadolsky, U. D. and B. Asquith. Quantifying the Impact of Human Immunodeficiency Virus-1 Escape From Cytotoxic T-Lymphocytes. *PLoS Comput Biol*, 6(11):e1000981, 2010.

[103] Kashiwagi, T., N. Hamada, J. Iwahashi, K. Hara, T. Ueda, *et al.* Emergence of new influenza A viruses which carry an escape mutation of the HLA-B27-restricted CTL epitope of NP in Japan. *Microbiol. Immunol.*, 44(10):867–870, 2000.

[104] van de Kassteele, J., J. van Eijkeren, and J. Wallinga. Efficient estimation of age-specific social contact rates between men and women. *Ann. Appl. Stat.*, 11(1):320–339, 2017.

[105] Kauffman, S. A. *The Origins of Order: Self Organization and Selection in Evolution*. Oxford University Press, 1993.

[106] Kawashima, Y., K. Pfafferott, J. Frater, P. Matthews, R. Payne, *et al.* Adaptation of HIV-1 to human leukocyte antigen class I. *Nature*, 458(7238):641–645, 2009.

[107] Kessinger, T. A., A. S. Perelson, and R. A. Neher. Inferring HIV escape rates from multi-locus genotype data. *Frontiers in T Cell Biology*, 4:252, 2013.

[108] Kiepiela, P., A. J. Leslie, I. Honeyborne, D. Ramduth, C. Thobakgale, *et al.* Dominant influence of HLA-B in mediating the potential co-evolution of HIV and HLA. *Nature*, 432(7018):769–775, 2004.

[109] Kiepiela, P., K. Ngumbela, C. Thobakgale, D. Ramduth, I. Honeyborne, *et al.* CD8+ T-cell responses to different HIV proteins have discordant associations with viral load. *Nature Medicine*, 13(1):46–53, 2007.

[110] Kinloch, N. N., D. R. MacMillan, A. Q. Le, L. A. Cotton, D. R. Bangsberg, *et al.* Population-Level Immune-Mediated Adaptation in HIV-1 Polymerase during the North American Epidemic. *J. Virol.*, 90(3):1244–1258, 2016.

[111] Koelle, K. and D. A. Rasmussen. The effects of a deleterious mutation load on patterns of influenza A/H3N2's antigenic evolution in humans. *Elife*, 4:e07361, 2015.

[112] Kotturi, M. F., I. Scott, T. Wolfe, B. Peters, J. Sidney, *et al.* Naive precursor frequencies and MHC binding rather than the degree of epitope diversity shape CD8+ T cell immunodominance. *J. Immunol.*, 181(3):2124–2133, 2008.

[113] Kouyos, R. D., G. E. Leventhal, T. Hinkley, M. Haddad, J. M. Whitcomb, *et al.* Exploring the complexity of the HIV-1 fitness landscape. *PLoS Genet.*, 8(3):e1002551, 2012.

[114] van der Kuyl, A. C., S. Jurriaans, G. Pollakis, M. Bakker, and M. Cornelissen. HIV RNA levels in transmission sources only weakly predict plasma viral load in recipients:. *AIDS*, 24(10):1607–1608, 2010.

[115] La Gruta, N., A. Kelso, L. E. Brown, W. Chen, D. C. Jackson, *et al.* Role of CD8(+) T-cell immunity in influenza infection: potential use in future vaccine development. *Expert Rev Respir Med*, 3(5):523–537, 2009.

[116] La Gruta, N. L. and S. J. Turner. T cell mediated immunity to influenza: mechanisms of viral control. *Trends Immunol.*, 35(8):396–402, 2014.

[117] Lalvani, A., R. Brookes, S. Hambleton, W. J. Britton, A. V. Hill, *et al.* Rapid effector function in CD8+ memory T cells. *J. Exp. Med.*, 186(6):859–865, 1997.

[118] Lang, G. I., D. P. Rice, M. J. Hickman, E. Sodergren, G. M. Weinstock, *et al.* Pervasive genetic hitchhiking and clonal interference in forty evolving yeast populations. *Nature*, 500(7464):571–574, 2013.

[119] Lapedes, A. and R. Farber. The geometry of shape space: application to influenza. *J. Theor. Biol.*, 212(1):57–69, 2001.

[120] Laurie, K. L., L. A. Carolan, D. Middleton, S. Lowther, A. Kelso, *et al.* Multiple infections with seasonal influenza A virus induce cross-protective immunity against A(H1N1) pandemic influenza virus in a ferret model. *J. Infect. Dis.*, 202(7):1011–1020, 2010.

[121] Leitner, T., B. Korber, M. Daniels, C. Calef, and B. Foley. HIV-1 subtype and circulating recombinant form (CRF) reference sequences. In Leitner, T., B. Foley, B. Hahn, P. Marx, F. McCutchan, *et al.*, editors, *HIV Sequence Compendium*, pages 41–48. Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, 2005.

[122] Leslie, A., D. Kavanagh, I. Honeyborne, K. Pfafferott, C. Edwards, *et al.* Transmission and accumulation of CTL escape variants drive negative associations between HIV polymorphisms and HLA. *J. Exp. Med.*, 201(6):891–902, 2005.

[123] Leslie, A., P. C. Matthews, J. Listgarten, J. M. Carlson, C. Kadie, *et al.* Additive contribution of HLA class I alleles in the immune control of HIV-1 infection. *J. Virol.*, 84(19):9879–9888, 2010.

[124] Leslie, A. J., K. J. Pfafferott, P. Chetty, R. Draenert, M. M. Addo, *et al.* HIV evolution: CTL escape mutation and reversion after transmission. *Nat. Med.*, 10(3):282–289, 2004.

[125] Lessler, J., S. Riley, J. M. Read, S. Wang, H. Zhu, *et al.* Evidence for antigenic seniority in influenza A (H3N2) antibody responses in southern China. *PLoS Pathog.*, 8(7):e1002802, 2012.

[126] Levin, B. R. and J. J. Bull. Short-sighted evolution and the virulence of pathogenic microorganisms. *Trends Microbiol.*, 2(3):76–81, 1994.

[127] Leviyang, S. Computational inference methods for selective sweeps arising in acute HIV infection. *Genetics*, 194(3):737–752, 2013.

[128] Lion, S. and J. A. J. Metz. Beyond R0 Maximisation: On Pathogen Evolution and Environmental Dimensions. *Trends Ecol. Evol. (Amst.)*, 2018.

[129] Little, S. J., A. R. McLean, C. A. Spina, D. D. Richman, and D. V. Havlir. Viral Dynamics of Acute HIV-1 Infection. *The Journal of Experimental Medicine*, 190(6):841–850, 1999.

[130] Liu, M. K., N. Hawkins, A. J. Ritchie, V. V. Ganusov, V. Whale, *et al.* Vertical T cell immunodominance and epitope entropy determine HIV-1 escape. *The Journal of Clinical Investigation*, 123(1):380–393, 2013.

[131] Liu, Y., J. P. McNevin, S. Holte, M. J. McElrath, and J. I. Mullins. Dynamics of Viral Evolution and CTL Responses in HIV-1 Infection. *PLoS ONE*, 6(1):e15639, 2011.

[132] Louie, R. H. Y., K. J. Kaczorowski, J. P. Barton, A. K. Chakraborty, and M. R. McKay. Fitness landscape of the human immunodeficiency virus envelope protein that is targeted by antibodies. *Proc. Natl. Acad. Sci. U.S.A.*, 115(4):E564–E573, 2018.

[133] Lowen, A. C., S. Mubareka, J. Steel, and P. Palese. Influenza virus transmission is dependent on relative humidity and temperature. *PLoS Pathog.*, 3(10):1470–1476, 2007.

[134] Lowen, A. C. and J. Steel. Roles of humidity and temperature in shaping influenza seasonality. *J. Virol.*, 88(14):7692–7695, 2014.

[135] Łuksza, M. and M. Lässig. A predictive fitness model for influenza. *Nature*, 507(7490):57–61, 2014.

[136] Lythgoe, K. A., F. Blanquart, L. Pellis, and C. Fraser. Large Variations in HIV-1 Viral Load Explained by Shifting-Mosaic Metapopulation Dynamics. *PLoS Biol.*, 14(10):e1002567, 2016.

[137] Lythgoe, K. A. and C. Fraser. New insights into the evolutionary rate of HIV-1 at the within-host and epidemiological levels. *Proc. Biol. Sci.*, 279(1741):3367–3375, 2012.

[138] Lythgoe, K. A., A. Gardner, O. G. Pybus, and J. Grove. Short-Sighted Virus Evolution and a Germline Hypothesis for Chronic Viral Infections. *Trends Microbiol.*, 25(5):336–348, 2017.

[139] Lythgoe, K. A., L. Pellis, and C. Fraser. Is HIV short-sighted? Insights from a multistrain nested model. *Evolution*, 67(10):2769–2782, 2013.

[140] Machkovech, H. M., T. Bedford, M. A. Suchard, and J. D. Bloom. Positive Selection in CD8+ T-Cell Epitopes of Influenza Virus Nucleoprotein Revealed by a Comparative Analysis of Human and Swine Viral Lineages. *J. Virol.*, 89(22):11275–11283, 2015.

[141] Mansky, L. M. and H. M. Temin. Lower in vivo mutation rate of human immunodeficiency virus type 1 than that predicted from the fidelity of purified reverse transcriptase. *J. Virol.*, 69(8):5087–5094, 1995.

[142] Markowitz, M., M. Louie, A. Hurley, E. Sun, M. Di Mascio, *et al.* A novel antiviral intervention results in more accurate assessment of human immunodeficiency virus type 1 replication dynamics and T-cell decay in vivo. *J. Virol.*, 77(8):5037–5038, 2003.

[143] McDonald, S. A., L. van Asten, W. van der Hoek, G. A. Donker, and J. Wallinga. The impact of national vaccination policy changes on influenza incidence in the Netherlands. *Influenza Other Respir Viruses*, 10(2):76–85, 2016.

[144] McDonald, S. A., M. van Boven, and J. Wallinga. An evidence synthesis approach to estimating the proportion of influenza among influenza-like illness patients. *Epidemiology*, 2017.

[145] McDonald, S. A., A. M. Presanis, D. De Angelis, W. van der Hoek, M. Hooiveld, *et al.* An evidence synthesis approach to estimating the incidence of seasonal influenza in the Netherlands. *Influenza Other Respir Viruses*, 8(1):33–41, 2014.

[146] McLaren, P. J. and M. Carrington. The impact of host genetic variation on infection with HIV-1. *Nat. Immunol.*, 16(6):577–583, 2015.

[147] McMichael, A. and P. Klenerman. HIV/AIDS. HLA leaves its footprints on HIV. *Science*, 296(5572):1410–1411, 2002.

[148] McMichael, A. J., F. M. Gotch, G. R. Noble, and P. A. Beare. Cytotoxic T-cell immunity to influenza. *N. Engl. J. Med.*, 309(1):13–17, 1983.

[149] Mellors, J. W., A. Munoz, J. V. Giorgi, J. B. Margolick, C. J. Tassoni, *et al.* Plasma Viral Load and CD4+ Lymphocytes as Prognostic Markers of HIV-1 Infection. *Annals of Internal Medicine*, 126(12):946–954, 1997.

[150] Meyer, D., V. R. C Aguiar, B. D. Bitarello, D. Y. C Brandt, and K. Nunes. A genomic perspective on HLA evolution. *Immunogenetics*, 70(1):5–27, 2018.

[151] Meyer, D., R. M. Singe, S. J. Mack, A. Lancaster, M. P. Nelson, *et al.* Single locus polymorphism of classical hla genes. In Hansen, J. A., editor, *Immunobiology of the Human MHC: Proceedings of the 13th International Histocompatibility Workshop and Conference, Volume I*, pages 653–704. IHWG Press, 2007.

[152] Mitov, V. and T. Stadler. A practical guide to estimating the heritability of pathogen traits. *Molecular Biology and Evolution*, 35(3):756–772, 2018.

[153] Monaco, D. C., D. A. Dilernia, A. Fiore-Gartland, T. Yu, J. L. Prince, *et al.* Balance between transmitted HLA preadapted and nonassociated polymorphisms is a major determinant of HIV-1 disease progression. *J. Exp. Med.*, 213(10):2049–2063, 2016.

[154] Moore, C. B., M. John, I. R. James, F. T. Christiansen, C. S. Witt, *et al.* Evidence of HIV-1 adaptation to HLA-restricted immune responses at a population level. *Science*, 296(5572):1439–1443, 2002.

[155] Moore, J., H. Ahmed, and R. Antia. High dimensional random walks can appear low dimensional: Application to influenza H3N2 evolution. *J. Theor. Biol.*, 447:56–64, 2018.

[156] Morens, D. M., J. K. Taubenberger, and A. S. Fauci. The persistent legacy of the 1918 influenza virus. *N. Engl. J. Med.*, 361(3):225–229, 2009.

[157] Morris, D. H., K. M. Gostic, S. Pompei, T. Bedford, M. ?uksza, *et al.* Predictive Modeling of Influenza Shows the Promise of Applied Evolutionary Biology. *Trends Microbiol.*, 26(2):102–118, 2018.

[158] Mossong, J., N. Hens, M. Jit, P. Beutels, K. Auranen, *et al.* Social contacts and mixing patterns relevant to the spread of infectious diseases. *PLoS Med.*, 5(3):e74, 2008.

[159] Mukherjee, S. and N. Chandra. Grouping of large populations into few CTL immune 'response-types' from influenza H1N1 genome analysis. *Clin Transl Immunology*, 3(8):e24, 2014.

[160] Müller, V., C. Fraser, and J. T. Herbeck. A Strong Case for Viral Genetic Factors in HIV Virulence. *Viruses*, 3(12):204–216, 2011.

[161] Müller, V., A. F. Maree, and R. J. De Boer. Small variations in multiple parameters account for wide variations in HIV-1 set-points: a novel modelling approach. *Proc. Biol. Sci.*, 268(1464):235–242, 2001.

[162] Mylius, S. D. and O. Diekmann. On evolutionarily stable life histories, optimization and the need to be specific about density dependence. *Oikos*, 74(2):218–224, 1995.

[163] Nagaraja, P., H. K. Alexander, S. Bonhoeffer, and N. M. Dixit. Influence of recombination on acquisition and reversion of immune escape and compensatory mutations in HIV-1. *Epidemics*, 14:11–25, 2016.

[164] Ndhlovu, Z. M., P. Kamya, N. Mewalal, H. N. Kl?verpris, T. Nkosi, *et al.* Magnitude and Kinetics of CD8(+) T Cell Activation during Hyperacute HIV Infection Impact Viral Set Point. *Immunity*, 43(3):591–604, 2015.

[165] Neher, R. A. and T. Bedford. nextflu: real-time tracking of seasonal influenza virus evolution in humans. *Bioinformatics*, 31(21):3546–3548, 2015.

[166] Neher, R. A., T. Bedford, R. S. Daniels, C. A. Russell, and B. I. Shraiman. Prediction, dynamics, and visualization of antigenic phenotypes of seasonal influenza viruses. *Proc. Natl. Acad. Sci. U.S.A.*, 113(12):E1701–1709, 2016.

[167] van Noort, S. P., R. Aguas, S. Ballesteros, and M. G. Gomes. The role of weather on the relation between influenza and influenza-like illness. *J. Theor. Biol.*, 298:131–137, 2012.

[168] van Noort, S. P., C. T. Codeco, C. E. Koppeschaar, M. van Ranst, D. Paolotti, *et al.* Ten-year performance of Influenzanet: ILI time series, risks, vaccine effects, and care-seeking behaviour. *Epidemics*, 13:28–36, 2015.

[169] Norris, J. R. *Markov Chains*. Cambridge University Press, 1998.

[170] Nowak, M. A. and C. R. Bangham. Population dynamics of immune responses to persistent viruses. *Science*, 272(5258):74–79, 1996.

[171] Nowak, M. A., R. M. May, R. E. Phillips, S. Rowland-Jones, D. G. Lalloo, *et al.* Antigenic oscillations and shifting immunodominance in HIV-1 infections. *Nature*, 375(6532):606–611, 1995.

[172] Pandit, A. and R. J. de Boer. Reliable reconstruction of HIV-1 whole genome haplotypes reveals clonal interference and genetic hitchhiking among immune escape variants. *Retrovirology*, 11:56, 2014.

[173] Payne, R., M. Muenchhoff, J. Mann, H. E. Roberts, P. Matthews, *et al.* Impact of HLA-driven HIV adaptation on virulence in populations of high HIV seroprevalence. *Proc. Natl. Acad. Sci. U.S.A.*, 111(50):E5393–5400, 2014.

[174] Perelson, A. S. and G. F. Oster. Theoretical studies of clonal selection: minimal antibody repertoire size and reliability of self-non-self discrimination. *J. Theor. Biol.*, 81(4):645–670, 1979.

[175] Phillips, A. N. Reduction of HIV concentration during acute infection: independence from a specific immune response. *Science (New York, N.Y.)*, 271(5248):497–499, 1996.

[176] Price, G. E., R. Ou, H. Jiang, L. Huang, and D. Moskophidis. Viral escape by selection of cytotoxic T cell-resistant variants in influenza A virus pneumonia. *J. Exp. Med.*, 191(11):1853–1867, 2000.

[177] Price, G. R. Selection and covariance. *Nature*, 227(5257):520–521, 1970.

[178] Prince, J. L., D. T. Claiborne, J. M. Carlson, M. Schaefer, T. Yu, *et al.* Role of Transmitted Gag CTL Polymorphisms in Defining Replicative Capacity and Early HIV-1 Pathogenesis. *PLoS Pathog*, 8(11):e1003041, 2012.

[179] Quinones-Parra, S., E. Grant, L. Loh, T. H. Nguyen, K. A. Campbell, *et al.* Preexisting CD8+ T-cell immunity to the H7N9 influenza A virus varies across ethnicities. *Proc. Natl. Acad. Sci. U.S.A.*, 111(3):1049–1054, 2014.

[180] Ramratnam, B., S. Bonhoeffer, J. Binley, A. Hurley, L. Zhang, *et al.* Rapid production and clearance of HIV-1 and hepatitis C virus assessed by large volume plasma apheresis. *Lancet*, 354(9192):1782–1785, 1999.

[181] Rao, X., A. I. C. A. F. Costa, D. v. Baarle, and C. Keşmir. A Comparative Study of HLA Binding Affinity and Ligand Diversity: Implications for Generating Immunodominant CD8+ T Cell Responses. *The Journal of Immunology*, 182(3):1526–1532, 2009.

[182] Rao, X., I. Hoof, A. I. C. A. F. Costa, D. v. Baarle, and C. Keşmir. HLA class I allele promiscuity revisited. *Immunogenetics*, 63(11):691–701, 2011.

[183] Ratmann, O., G. Donker, A. Meijer, C. Fraser, and K. Koelle. Phylodynamic inference and model assessment with approximate bayesian computation: influenza as a case study. *PLoS Comput. Biol.*, 8(12):e1002835, 2012.

[184] Refsland, E. W., J. F. Hultquist, E. M. Luengas, T. Ikeda, N. M. Shaban, *et al.* Natural polymorphisms in human APOBEC3H and HIV-1 Vif combine in primary T lymphocytes to affect viral G-to-A mutation levels and infectivity. *PLoS Genet.*, 10(11):e1004761, 2014.

[185] Regoes, R. R., R. Antia, D. A. Garber, G. Silvestri, M. B. Feinberg, *et al.* Roles of Target Cells and Virus-Specific Cellular Immunity in Primary Simian Immunodeficiency Virus Infection. *Journal of Virology*, 78(9):4866–4875, 2004.

[186] Regoes, R. R., P. J. McLaren, M. Battegay, E. Bernasconi, A. Calmy, *et al.* Disentangling human tolerance and resistance against HIV. *PLoS Biol.*, 12(9):e1001951, 2014.

[187] Rimmelzwaan, G. F., A. C. Boon, J. T. Voeten, E. G. Berkhoff, R. A. Fouchier, *et al.* Sequence variation in the influenza A virus nucleoprotein associated with escape from cytotoxic T lymphocytes. *Virus Res.*, 103(1-2):97–100, 2004.

[188] Roberts, H. E., J. Hurst, N. Robinson, H. Brown, P. Flanagan, *et al.* Structured observations reveal slow HIV-1 CTL escape. *PLoS Genet.*, 11(2):e1004914, 2015.

[189] Robinson, J., L. A. Guethlein, N. Cereb, S. Y. Yang, P. J. Norman, *et al.* Distinguishing functional polymorphism from random variation in the sequences of >10,000 HLA-A, -B and -C alleles. *PLoS Genet.*, 13(6):e1006862, 2017.

[190] Robinson, J., J. A. Halliwell, H. McWilliam, R. Lopez, and S. G. Marsh. IPD–the Immuno Polymorphism Database. *Nucleic Acids Res.*, 41(Database issue):D1234–1240, 2013.

[191] Rodriguez, B., A. K. Sethi, V. K. Cheruvu, W. Mackay, R. J. Bosch, *et al.* Predictive value of plasma HIV RNA level on rate of CD4 T-cell decline in untreated HIV infection. *JAMA*, 296(12):1498–1506, 2006.

[192] Rosenthal, J. S. Optimal Proposal Distributions and Adaptive MCMC. In Brooks, S., A. Gelman, G. L. Jones, and X.-L. Meng, editors, *Handbook of Markov Chain Monte Carlo*, chapter 4, pages 93–112. Chapman and Hall/CRC, Boca Raton, 2011.

[193] Rosseel, Y. lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software*, 48(2):1–36, 2012.

[194] Saez-Cirion, A., C. Bacchus, L. Hocqueloux, V. Avettand-Fenoel, I. Girault, *et al.* Post-treatment HIV-1 controllers with a long-term virological remission after the interruption of early initiated antiretroviral therapy ANRS VISCONTI Study. *PLoS Pathog.*, 9(3):e1003211, 2013.

[195] Sallusto, F., J. Geginat, and A. Lanzavecchia. Central memory and effector memory T cell subsets: function, generation, and maintenance. *Annu. Rev. Immunol.*, 22:745–763, 2004.

[196] van de Sandt, C. E., J. H. Kreijtz, and G. F. Rimmelzwaan. Evasion of influenza A viruses from innate and adaptive immune responses. *Viruses*, 4(9):1438–1476, 2012.

[197] van de Sandt, C. E., M. L. Hillaire, M. M. Geelhoed-Mieras, A. D. Osterhaus, R. A. Fouchier, *et al.* Human Influenza A Virus-Specific CD8+ T-Cell Response Is Long-lived. *J. Infect. Dis.*, 212(1):81–85, 2015.

[198] Sanjuán, R., M. R. Nebot, J. B. Peris, and J. Alcamí. Immune Activation Promotes Evolutionary Conservation of T-Cell Epitopes in HIV-1. *PLoS Biol*, 11(4):e1001523, 2013.

[199] Schellens, I. M., M. Navis, H. W. van Deutekom, B. Boeser-Nunnink, B. Berkhout, *et al.* Loss of HIV-1-derived cytotoxic T lymphocyte epitopes restricted by protective HLA-B alleles during the HIV-1 epidemic:. *AIDS*, 25(14):1691–1700, 2011.

[200] Schmid, B., C. Kęsmir, and R. de Boer. Quantifying how MHC polymorphism prevents pathogens from adapting to the antigen presentation pathway. *Epidemics*, 2(3):99–108, 2010.

[201] Schmitz, J. E., M. J. Kuroda, S. Santra, V. G. Sasseville, M. A. Simon, *et al.* Control of Viremia in Simian Immunodeficiency Virus Infection by CD8+ Lymphocytes. *Science*, 283(5403):857–860, 1999.

[202] Sellke, T. On the Asymptotic Distribution of the Size of a Stochastic Epidemic. *Journal of Applied Probability*, 20(2):390–394, 1983.

[203] Sette, A. and J. Sidney. Nine major HLA class I supertypes account for the vast preponderance of HLA-A and -B polymorphism. *Immunogenetics*, 50(3-4):201–212, 1999.

[204] Sharp, P. M., E. Bailes, R. R. Chaudhuri, C. M. Rodenburg, M. O. Santiago, *et al.* The origins of acquired immune deficiency syndrome viruses: where and when? *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 356(1410):867–876, 2001.

[205] Shirreff, G., S. Alizon, A. Cori, H. F. Gunthard, O. Laeyendecker, *et al.* How effectively can HIV phylogenies be used to measure heritability? *Evol Med Public Health*, 2013(1):209–224, 2013.

[206] Shirreff, G., L. Pellis, O. Laeyendecker, and C. Fraser. Transmission selects for HIV-1 strains of intermediate virulence: a modelling approach. *PLoS Comput. Biol.*, 7(10):e1002185, 2011.

[207] Shubin, M., A. Lebedev, O. Lyytikainen, and K. Auranen. Revealing the True Incidence of Pandemic A(H1N1)pdm09 Influenza in Finland during the First Two Seasons - An Analysis Based on a Dynamic Transmission Model. *PLoS Comput. Biol.*, 12(3):e1004803, 2016.

[208] Sidney, J., B. Peters, N. Frahm, C. Brander, and A. Sette. HLA class I supertypes: a revised and updated classification. *BMC Immunol.*, 9:1, 2008.

[209] Sievers, F., A. Wilm, D. Dineen, T. J. Gibson, K. Karplus, *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.*, 7:539, 2011.

[210] Smith, D. J., A. S. Lapedes, J. C. de Jong, T. M. Bestebroer, G. F. Rimmelzwaan, *et al.* Mapping the antigenic and genetic evolution of influenza virus. *Science*, 305(5682):371–376, 2004.

[211] Smith, D. R. and N. Mideo. Modelling the evolution of HIV-1 virulence in response to imperfect therapy and prophylaxis. *Evol Appl*, 10(3):297–309, 2017.

[212] Spruijt, I. T., M. M. de Lange, F. Dijkstra, G. A. Donker, and W. van der Hoek. Long-Term Correlation between Influenza Vaccination Coverage and Incidence of Influenza-Like Illness in 14 European Countries. *PLoS ONE*, 11(9):e0163508, 2016.

[213] Sridhar, S. Heterosubtypic T-Cell Immunity to Influenza in Humans: Challenges for Universal T-Cell Influenza Vaccines. *Front Immunol*, 7:195, 2016.

[214] Sridhar, S., S. Begom, A. Bermingham, K. Hoschler, W. Adamson, *et al.* Cellular immune correlates of protection against symptomatic pandemic influenza. *Nat. Med.*, 19(10):1305–1312, 2013.

[215] Storey, J. D. and R. Tibshirani. Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. U.S.A.*, 100(16):9440–9445, 2003.

[216] Strelkowa, N. and M. Lässig. Clonal interference in the evolution of influenza. *Genetics*, 192(2):671–682, 2012.

[217] Tang, J., S. Tang, E. Lobashevsky, I. Zulu, G. Aldrovandi, *et al.* HLA Allele Sharing and HIV Type 1 Viremia in Seroconverting Zambians with Known Transmitting Partners. *AIDS Research and Human Retroviruses*, 20(1):19–25, 2004.

[218] Teirlinck, A. C., L. Van Asten, P. S. Brandsema, F. Dijkstra, G. A. Donker, *et al.* Annual report surveillance of influenza and other respiratory infections in the netherlands: winter 2014/2015. www.rivm.nl, 2015. RIVM report number 2015-0042.

[219] The International HIV Controllers Study. The Major Genetic Determinants of HIV-1 Control Affect HLA Class I Peptide Presentation. *Science*, 330(6010):1551–1557, 2010.

[220] Thomas, P. G., R. Keating, D. J. Hulse-Post, and P. C. Doherty. Cell-mediated protection in influenza infection. *Emerging Infect. Dis.*, 12(1):48–54, 2006.

[221] Thyagarajan, B. and J. D. Bloom. The inherent mutational tolerance and antigenic evolvability of influenza hemagglutinin. *Elife*, 3, 2014.

[222] Valkenburg, S. A., S. Quinones-Parra, S. Gras, N. Komadina, J. McVernon, *et al.* Acute emergence and reversion of influenza A virus quasispecies within CD8+ T cell antigenic peptides. *Nat Commun*, 4:2663, 2013.

[223] Vidya Vijayan, K. K., K. P. Karthigeyan, S. P. Tripathi, and L. E. Hanna. Pathophysiology of CD4+ T-Cell Depletion in HIV-1 and HIV-2 Infections. *Front Immunol*, 8:580, 2017.

[224] Vita, R., L. Zarebski, J. A. Greenbaum, H. Emami, I. Hoof, *et al.* The immune epitope database 2.0. *Nucleic Acids Res.*, 38(Database issue):D854–862, 2010.

[225] Volz, E. M., E. Ionides, E. O. Romero-Severson, M. G. Brandt, E. Mokotoff, *et al.* HIV-1 transmission during early infection in men who have sex with men: a phylodynamic analysis. *PLoS Med.*, 10(12):e1001568; discussion e1001568, 2013.

[226] Wallinga, J., P. Teunis, and M. Kretzschmar. Using data on social contacts to estimate age-specific transmission parameters for respiratory-spread infectious agents. *Am. J. Epidemiol.*, 164(10):936–944, 2006.

[227] Wang, S., P. Hottz, M. Schechter, and L. Rong. Modeling the Slow CD4+ T Cell Decline in HIV-Infected Individuals. *PLoS Comput. Biol.*, 11(12):e1004665, 2015.

[228] Watanabe, S. Asymptotic Equivalence of Bayes Cross Validation and Widely Applicable Information Criterion in Singular Learning Theory. *Journal of Machine Learning Research*, 11:3571–3594, 2010.

[229] Watanabe, S. A Widely Applicable Bayesian Information Criterion. *Journal of Machine Learning Research*, 14:867–897, 2013.

[230] Wawer, M. J., R. H. Gray, N. K. Sewankambo, D. Serwadda, X. Li, *et al.* Rates of HIV-1 Transmission per Coital Act, by Stage of HIV-1 Infection, in Rakai, Uganda. *Journal of Infectious Diseases*, 191(9):1403–1409, 2005.

[231] Webster, R. G., W. J. Bean, O. T. Gorman, T. M. Chambers, and Y. Kawaoka. Evolution and ecology of influenza A viruses. *Microbiol. Rev.*, 56(1):152–179, 1992.

[232] Weinfurter, J. T., K. Brunner, S. V. Capuano, C. Li, K. W. Broman, *et al.* Cross-reactive T cells are involved in rapid clearance of 2009 pandemic H1N1 influenza virus in nonhuman primates. *PLoS Pathog.*, 7(11):e1002381, 2011.

[233] Westgeest, K. B., M. de Graaf, M. Fourment, T. M. Bestebroer, R. van Beek, *et al.* Genetic evolution of the neuraminidase of influenza A (H3N2) viruses from 1968 to 2009 and its correspondence to haemagglutinin evolution. *J. Gen. Virol.*, 93(Pt 9):1996–2007, 2012.

[234] Wikramaratna, P. S., J. Lourenco, P. Klenerman, O. G. Pybus, and S. Gupta. Effects of neutralizing antibodies on escape from CD8+ T-cell responses in HIV-1 infection. *Philos. Trans. R. Soc. Lond., B, Biol. Sci.*, 370(1675), 2015.

[235] Wikramaratna, P. S., M. Sandeman, M. Recker, and S. Gupta. The antigenic evolution of influenza: drift or thrift? *Philos. Trans. R. Soc. Lond., B, Biol. Sci.*, 368(1614):20120200, 2013.

[236] Wilkinson, T. M., C. K. Li, C. S. Chui, A. K. Huang, M. Perkins, *et al.* Preexisting influenza-specific CD4+ T cells correlate with disease protection against influenza challenge in humans. *Nat. Med.*, 18(2):274–280, 2012.

[237] de Wolf, F., I. Spijkerman, P. T. Schellekens, M. Langendam, C. Kuiken, *et al.* AIDS prognosis based on HIV-1 RNA, CD4+ T-cell count and function: markers with reciprocal predictive value over time after seroconversion. *AIDS*, 11(15):1799–1806, 1997.

[238] Woolthuis, R. G., C. H. van Dorp, C. Keşmir, R. J. de Boer, and M. van Boven. Long-term adaptation of the influenza A virus by escaping cytotoxic T-cell recognition. *Sci Rep*, 6:33334, 2016.

[239] Worobey, M., M. Gemmel, D. E. Teuwen, T. Haselkorn, K. Kunstman, *et al.* Direct evidence of extensive diversity of HIV-1 in Kinshasa by 1960. *Nature*, 455(7213):661–664, 2008.

[240] Worobey, M., G. Z. Han, and A. Rambaut. A synchronized global sweep of the internal genes of modern avian influenza virus. *Nature*, 508(7495):254–257, 2014.

[241] Xia, Y., J. R. Gog, and B. T. Grenfell. Semiparametric estimation of the duration of immunity from infectious disease time series: influenza as a case-study. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54(3):659–672, 2005.

[242] Yang, W., M. Lipsitch, and J. Shaman. Inference of seasonal and pandemic influenza transmission dynamics. *Proc. Natl. Acad. Sci. U.S.A.*, 112(9):2723–2728, 2015.

[243] Yang, Z. and C. E. Rodriguez. Searching for efficient Markov chain Monte Carlo proposal kernels. *Proc. Natl. Acad. Sci. U.S.A.*, 110(48):19307–19312, 2013.

[244] Yates, A., J. Stark, N. Klein, R. Antia, and R. Callard. Understanding the slow depletion of memory CD4+ T cells in HIV infection. *PLoS Med.*, 4(5):e177, 2007.

[245] Yue, L., H. A. Prentice, P. Farmer, W. Song, D. He, *et al.* Cumulative Impact of Host and Viral Factors on HIV-1 Viral-Load Control during Early Infection. *Journal of Virology*, 87(2):708–715, 2013.