Exploring the effects of immunity and life history on the dynamics of an endogenous retrovirus

R. K. Kanda¹, M. Tristem² and T. Coulson¹

¹Department of Zoology, University of Oxford, Tinbergen Building, South Parks Road, Oxford OX1 3PS, UK
²Department of Life Sciences, Imperial College London, Silwood Park, Ascot, Berkshire SL5 7PY, UK

Mammalian DNA is littered with the signatures of past retroviral infections. For example, at least 8% of the human genome can be attributed to endogenous retroviruses (ERVs). We take a single-locus approach to develop a simple susceptible–infected–recovered model to investigate the circumstances under which a disease-causing retrovirus can become incorporated into the host genome and spread through the host population if it were to confer an immunological advantage. In the absence of any fitness benefit provided by the long terminal repeat (LTR), we conclude that signatures of ERVs are likely to go to fixation within a population when the probability of evolving cellular/humoral immunity to a related exogenous version of the virus is extremely small. We extend this model to examine whether changing the speed of the host life history influences the likelihood that an exogenous retrovirus will incorporate and spread to fixation. Our results reveal the parameter space under which incorporation of exogenous retroviruses into a host genome may be beneficial to the host. In our final model, we find that the likelihood of an LTR reaching fixation in a host population is not strongly affected by host life history.

1. Introduction

Mammalian genomes are littered with the signatures of ancient retroviral infections. Endogenous retroviruses (ERVs) are formed by the successful incorporation of an exogenous retrovirus into the germline and its subsequent vertical transmission. ERVs have been described from a wide range of vertebrate genomes, including mammals, fish, birds, reptiles and amphibians [1]. They constitute a significant proportion of host genomes: in humans, ERVs represent 8% of the genome and over 10% of the genome in mice [2,3]. Typically, an ERV consists of three genes (gag, pol and env) and two flanking non-coding long terminal repeats (LTRs), which are identical at the time of integration. Over time, these insertions will have accumulated mutations and deletions at the same rate as the mutation rate of the host genome [4], rendering them non-functional. ERVs may also be inactivated by recombinational deletion between the two flanking LTRs, which removes the internal coding region, leaving a solo LTR. Solo LTRs are 10–100 times more numerous than their full-length counterparts [5], and many of these insertions are fixed in the host population. In various mammal species, there are a few examples of intact, evolutionarily young ERVs that are polymorphic (presence/absence) in their host population [6–11]; whether this is persisting polymorphism maintained through various evolutionary forces or actual active ERVs remains to be established. However, some ERVs do appear to be replication competent [10,12–16].

Although most insertions are inactive remnants subject to forces such as genetic drift, some have maintained open reading frames (ORFs), exhibit low \( d_{s}/d_{S} \) ratios (non-synonymous–synonymous mutations), or have fewer stop codons than expected under neutrality, indicating purifying selection [17,18] and therefore functional constraint. There are numerous examples, in a range of species, where a retroviral insertion has been selected for as it...
confers a fitness advantage on the host, perhaps through co-option of a retroviral gene or through gene regulation
[19–22]. There are also instances involving the host using the promoters in the LTRs of retroviruses to regulate
dexpression of a host gene close to the point of retroviral
insertion [21,23–25]. However, in most locations, these
insertions have no known function. The increasing avail-
bility of genomic data has provided unique opportunities
to study retroviral insertions, and recent advances have
revealed that a number of ERVs have been co-opted as cel-
lular genes which act as part of the immune system of the
host, inhibiting viral infection [26]. This is a remarkable
immune strategy in the evolutionary arms race between
viruses and their hosts: the integration of viruses for use
against themselves.

These observations raise some important questions:
under what circumstances do ERVs successfully spread
through host populations?Conventionally, it is thought that
once inserted into the genome the insertions we see have
drifted to fixation, but are there other mechanisms via
which an active ERV could provide a sufficient fitness advan-
tage to spread through a population, before losing its
apparent advantage when it reaches fixation, such that it
can be excised from the human genome? One plausible
suggestion is that for some ERVs, incorporation into the
genome provides the host with some immunity from related
exogenous retroviruses [27,28] (described as endogenous viral element derived immunity, EDI [26]). There are a
number of examples of ERVs conferring some immunity through a
variety of mechanisms [9,29,30]. Modelling provides a route
to exploring this possibility.

Compartmental models offer a useful tool to examine
both the ecological and evolutionary dynamics of ERVs
[31]. Such models are widely used to study the dynamics of
a range of pathogens [32], but have not been previously
used to study the consequences of ERV ‘epidemics’ on host
dynamics. Compartmental models structure a population
by the pathogen status of the host. The simplest models
have two states: those susceptible to infection (S) and those
infected (I) with the pathogen. More complicated models
include a recovered class (R) [31]. We develop a range of
models that can include multiple infected and recovered
states, and use these models to determine the conditions
under which incorporation of retroviruses into the host
genome benefits both the pathogen and the host.

We also investigate the impact of host life history on the
probability that an LTR reaches fixation at an arbitrary,
unspecified location in the host genome. The immune sys-
tems of short- and long-lived species can differ in many
important ways, some of which may influence the fitness of
a strategy that involves incorporation of an exogenous retro-
virus into the host genome. There is evidence to suggest that
immune defences may vary with the life history of the host
[33]—that different immune response strategies may be
used to a greater or lesser extent than others, depending on
certain life-history traits. Mouse genomes, in particular,
have many active ERVs, which contrasts strikingly with the
human genome, where all ERVs are nearly extinct, with
the possible exception of HERV-K [6], potentially suggesting
that there may be some aspects of the host, perhaps overall
exposure to retroviruses, that influences whether ERVs
remain active or become extinct. We extend the model to
examine whether changing the speed of the host life history
influences the likelihood that an exogenous retrovirus will
incorporate and its inactive signature spread to fixation.

2. Modelling philosophy

Our aim is to describe the initial invasion of a genome by a re-
trovirus. We develop a discrete time ‘SIR’ model that we can
use to examine the proportion of the population in each compo-
artment after an infection has run its course. The models we
construct consist of compartments, with parameters describ-
ing the rates of movement of individuals between classes
(figure 1) [32]. We develop a range of models of increasing
complexity, with the final model containing five compart-
ments. This allowed us a detailed understanding of the
predicted dynamics, even though we were unable to derive
exact analytical solutions to the more complex models. Our
use of compartmental models is novel as we are not interested
per se in the dynamics of the infection; rather we are interested
in the stable population structure after the epidemic has run
its course, and particularly in the proportion of the popula-
tion in the recovered state expressing a signature of the
retroviral infection.

Before a population is exposed to an exogenous retro-
virus, all individuals are in the susceptible compartment

![Figure 1](http://rstb.royalsocietypublishing.org/)

**Figure 1.** Graphical representation of (a) the SI model, (b) the SIR model, (c) the SIR\_R\_LTR model with \( I \) newborns entering the \( S \) compartment and recovered \( I \) newborns entering the \( S \) compartment, and (d) the SIR\_R\_LTR model with \( I \) newborns entering the \( I \) compartment and recovered \( I \) new-
borns entering the \( I \) compartment. The circles represent compartments; the
arrows represent transition rates between the compartments.
(S)—this means they are susceptible to becoming infected with the exogenous retrovirus. If they encounter an infected individual, then they have a probability of becoming infected. Once individuals are infected, then they can move into the infected with the exogenous virus compartment (I_X). Offspring born to individuals in this compartment are susceptible to infection, and they enter the susceptible compartment. Our first model is an SI model, consisting of just these two compartments (figure 1a). This model is useful because it is analytically tractable.

In our next model, we incorporate a second infected compartment for individuals that have incorporated the exogenous retrovirus into their germline (I_E). Individuals enter this class as offspring of parents that have incorporated the exogenous virus into the germline. These offspring have inherited the endogenous retroviral insertion, which is still producing infectious viral particles [34,35]. Offspring born of parents in this class will remain in the I_E compartment unless mutation deactivates the functional ERV, or recombinational deletion removes the ERV from their genome, in which case these individuals enter the susceptible compartment. For individuals in the I_E class, any such immunity provided by endogenization would be inherited. This model is an SI_E model. We were unable to obtain any useful analytical solutions to this model, but its analysis is useful as it helps us identify the range of conditions under which a large proportion of a population ends up with incorporated retroviruses at a particular locus in their genome. Consequently, modelling the dynamics as whether incorporation occurs at a particular site in the genome allows us to understand the processes that can lead to an endogenous retroviral signature spreading through a population. We analyse this model graphically over the complete range of possible parameter values for a single, approximately human, life history.

There is another route to leaving the I_E and I_X compartments: through acquiring immunity to either the endogenous or exogenous retroviruses, and entering respectively the recovered with, and recovered without, an endogenous retroviral signature in their DNA (R_E and R_X, figure 1c). Individuals in the R_N class will have mounted a successful immune response to the exogenous retroviral infection, causing inactivation or clearance of the exogenous virus [36]. Consequently, their offspring will have no inherited immunity, and they return to the susceptible class. By the time individuals leave the I_E compartment, and the ERV insertion has undergone recombinational deletion to produce the solo LTR (i.e. individuals enter the R_LTR class), we assume the threat of the exogenous retrovirus has passed. Whatever immunity was offered by the ERV insertion is no longer required and thus the selection pressure on retaining the full-length ERV insertion is removed. Individuals cannot leave this recovered class. This class of model is an absorbing Markov chain. Because we were unable to obtain analytical solutions to the SI_E model, we were unable to analytically analyse this model either. We consequently provide numerical results.

3. The models

(a) A susceptible—infected model

We start with the simple SI model (figure 1a) of the dynamics of an exogenous retrovirus to which immunity cannot occur [37]. Susceptible individuals move into the infected state as a function of the proportion of the population in the susceptible and infected compartment and an infection rate, β. This simple model can be written

\[ S(t + 1) = \left[ \phi_S + \tau \right] S(t) + \pi_X I_X(t) - \beta S(t) I_X(t) / S(t) + I_X(t) \]

and

\[ I_X(t + 1) = \phi_S \phi_X I_X(t) + \beta S(t) I_X(t) / S(t) + I_X(t) \]

(3.1)

where \( t \) represents time, \( \tau \) is the birth rate which does not differ between compartments, \( \phi_S \) is the survival rate of susceptible individuals and \( \phi_X \) is the survival rate of individuals infected with the exogenous retrovirus. Consequently, \( 1 - \phi_X \) is the mortality rate imposed by the exogenous virus. Individuals born to infected individuals enter the susceptible compartment: \( N(t) = S(t) + I_X(t) \).

We are interested in the proportion of the population in each of the two compartments. It is straightforward to solve this equation analytically, which reveals that the ratio of susceptible to infected individuals is

\[ S \over S_X = \phi_S(\phi_X - 1) + \beta \over \tau - 1 \]  

(3.2)

And the proportion of the population in the infected compartment is

\[ I_X \over S_X = 1 - \phi_S(\phi_X - 1) + \beta \]  

(3.3)

(b) Adding a second infected compartment

From this starting point, we can introduce a little more complexity into the model by introducing a second infected compartment (I_E) for individuals infected that have incorporated the virus into their genomes. For individuals to enter the I_E compartment, they must pass through the I_X compartment. We distinguish between the two infected states, and with parameters associated with each compartments, using the subscripts X (for exogenous) and E (for endogenous; figure 1b).

The SI_E model is of the form

\[ S(t + 1) = \left( \phi_S + \tau \right) S(t) + \pi_X I_X(t) - \beta_S S(t) I_X(t) / N(t) + \] 

\[ + (1 - \gamma) \pi_X I_X(t) + \alpha I_E(t) \]

\[ I_X(t + 1) = \left( \phi_S \phi_X I_X(t) + \beta_S S(t) I_X(t) / N(t) + \beta_E S(t) I_E(t) / N(t) \right) \]

and

\[ I_E(t + 1) = \left( \phi_S \phi_X + \tau I_E(t) + \gamma I_X(t) - \alpha I_E(t) \right) \]

(3.4)

Here, the new parameters are \( \chi \), the rate at which exogenous retroviruses incorporate into the host genome and \( \alpha \), the rate at which functioning ERVs are lost from the genome through mutation or recombinational deletion. We assume that individuals in the I_E compartment are less infectious than those in the I_X compartment (\( \beta_E \) < \( \beta_X \)) as exogenous retroviruses are typically more virulent than their endogenous counterparts. \( 1 - \phi_X \) is the mortality rate of individuals in compartment I_X. \( N(t) = S(t) + I_X(t) + I_E(t) \).

(c) Incorporating recovered compartments

We now expand the SI_E model to incorporate two recovered classes to derive an SIR model (specifically an SI_E_R_SLTR model).
model; figure 1c). Movement between the five compartments can mathematically be described with the following five equations:

\[
S(t+1) = (\phi_s + \tau)S(t) - \frac{\beta_c S(t)I_X(t)}{N(t)} - \frac{\beta_b S(t)I_L(t)}{N(t)} + (1 - \gamma)\tau I_X(t) + \alpha \tau I_L(t) + \tau R_N(t) \\
I_X(t+1) = (\phi_s \phi_b + \gamma)I_X(t) + \frac{\beta_c S(t)I_X(t)}{N(t)} + \frac{\beta_b S(t)I_L(t)}{N(t)} - \theta I_X(t) \\
I_L(t+1) = (\phi_s \phi_b + \gamma)I_L(t) + \frac{\beta_c S(t)I_L(t)}{N(t)} + \frac{\beta_b S(t)I_X(t)}{N(t)} - \gamma \theta I_L(t) - \alpha \tau I_E(t) + \alpha \tau I_L(t) \\
R_N(t+1) = \phi_b R_N(t) + \theta I_X(t) \\
and \ R_{LTR}(t+1) = (\phi_{LTR} + \gamma)R_{LTR}(t) + \theta \alpha \tau I_L(t) \\
\]

where \(\theta\) is the rate at which immunity occurs in the exogenous compartment, \(\gamma\) is the rate at which immunity occurs in the endogenous compartment, and \(N(t) = S(t) + I_X(t) + I_L(t) + R_S(t) + R_{LTR}(t)\).

Our initial simulations indicated that the model structure described above was unrealistic for faster life histories in that LTRs did not reach high frequencies within the population—there are many examples of species with faster life histories which display signatures of retroviral infection [3]. We therefore modified the model to make it more biologically realistic, and by relaxing some of the assumptions of the original \(S/I_L/R_N/R_{LTR}\) model, reduced the flow of individuals from right to left (figure 1d). Vertical transmission of retroviruses from mother to offspring is quite likely [36,38], and we altered the model to reflect this. Offspring born to infected mothers \(I_X\) now enter the \(I_X\) compartment, and offspring born to mothers with a functioning version of the endogenous retrovirus \(I_L\) that have experienced mutation or recombinational deletion, before immunity (EDI) has arisen, enter the \(I_X\) compartment. The modified equations are

\[
S(t+1) = (\phi_s + \tau)S(t) - \frac{\beta_c S(t)I_X(t)}{N(t)} - \frac{\beta_b S(t)I_L(t)}{N(t)} + \gamma \theta I_X(t) + \alpha \tau I_L(t) \\
I_X(t+1) = (\phi_s \phi_b + \gamma)I_X(t) + \frac{\beta_c S(t)I_X(t)}{N(t)} + \frac{\beta_b S(t)I_L(t)}{N(t)} - \theta I_X(t) - \gamma \theta I_L(t) - \alpha \tau I_E(t) \\
I_L(t+1) = (\phi_s \phi_b + \gamma)I_L(t) + \frac{\beta_c S(t)I_L(t)}{N(t)} + \frac{\beta_b S(t)I_X(t)}{N(t)} - \gamma \theta I_L(t) - \alpha \tau I_E(t) - \gamma \theta I_L(t) \\
R_N(t+1) = \phi_b R_N(t) + \theta I_X(t) \\
and \ R_{LTR}(t+1) = (\phi_{LTR} + \gamma)R_{LTR}(t) + \theta \alpha \tau I_L(t) \\
\]

(d) Parameter values for the model

The survival and fertility rates indirectly influence the rate of movement between all compartments—these values also determine the life history of the species [39]. Large survival rate parameters and small fertility rate parameters correspond to a species with a slow life history, whereas large fertility rate parameters and small survival rate parameters correspond to a species with a faster life history. The model can consequently be used to examine how life history influences the probability of incorporation of an exogenous retrovirus into the host genome.

The maximum population growth rate the population can achieve occurs when all individuals are in the \(S\) and \(R_{LTR}\) classes. When this occurs, the population growth rate is \(\phi_s + \tau\). We set this to be greater than unity to prevent extinction. For all models and simulations we report, we constrain \(\phi_s + \tau = 1.016\). We chose this constraint as extinction did not occur in our simulations for this value, but could occur if the value was smaller.

We alter the life history of the species by changing the values of \(\phi_s\) and \(\tau\) such that their sum always equals 1.016. As \(\tau\) gets larger, the life history speeds up, and as \(\phi_s\) increases the life history slows down. We vary \(\phi_s\) and \(\tau\) in increments of 0.01. For each life history, we then independently vary values of all other parameters between zero and unity (they are all rates). Simulations are run for 5000 time steps before we record the proportion of the population in the \(R_{LTR}\) compartment.

In our initial model, and in the absence of any published information on mortality rates attributable to ERVs, we conservatively impose a mortality rate of 0.03 attributable to the exogenous virus (1 - \(\phi_s\)), and a rate of 0.015 attributable to the endogenous virus (1 - \(\phi_b\)). We make the assumption that the mortality rate of the \(I_L\) group is less than that of the \(I_X\) compartment, because exogenous viruses typically kill host cells they infect, whereas ERVs do not. We impose a survival cost of incorporation of the virus into host DNA, because the host continues to produce virions. The \(I_X\) compartment is likely to produce more virions than the \(I_L\) group for the same reason that they have different mortality rates. For this reason, we also assume \(I_X\) individuals are more infectious; we set \(\beta_b = 0.4\) and \(\beta_s = 0.5\) in our initial model.

There is little available information about values for the other parameters in the literature. From one of the most studied genomes, the human genome, the rate of loss of an active ERV at a point in the genome via mutation or recombinational deletion is very low (in humans, the latest estimates of mutation rates are around 1.2 × 10^{-8} per site [40,41], and rates of recombinational deletion for HERVs are estimated at 4.3 × 10^{-7} to 8 × 10^{-5} [42]). Little is known about the rates of integration. For one of the youngest HERV families, HERV-K, many insertions are shared with chimpanzees. However, a number of HERV-K insertions are human specific (approx. 130; calculated from the reference sequence and available literature), having integrated after the divergence of humans and chimpanzee (approx. 7 Ma) [43]. These are clearly those that are assumed to have drifted to fixation—many others insertions would have been lost (due to drift). Assuming a generation time of 20 years and effective population size of 10,000, we can calculate that the probability of a new retroviral insert per genome per generation is 3.7 × 10^{-4}—again quite low. For our model, the key issue is that the rate of insertion per genome is greater than the rate of loss. We consequently assume \(\alpha < \gamma\). The specific values of alpha and gamma determine the rates of spread of the epidemic through the population, but not the equilibrium outcome.

Little is known about the rates at which immunity would arise and so we arbitrarily set \(\theta = 0.05\) and \(\theta' = 0.01\). Some of these values may seem rather high. However, analysis of the model reveals that the dynamics it predicts are relatively constant across a wide range of parameter values (see §4) and that one of our key results depends upon the ratio of \(\theta\) to \(\theta'\) (§4). In other words, the absolute parameter values of \(\theta\), \(\theta'\) and \(\alpha\) are relatively unimportant; it is their relative values that determine the dynamics. As the values of \(\theta\), \(\theta'\) and \(\alpha\) get closer to
Figure 2. The compete dynamics of the SI/E model represented graphically. Colours represent the proportion of the equilibrium population in the I,E compartment, with red representing a proportion close to zero and light yellow representing a proportion close to unity. Each of the x-axis on each image plot represents the infection rate by individuals infected with the endogenous retrovirus, and the y-axis represents the infection rate by individuals infected with the exogenous retrovirus. Both axes span the range 0 – 1. Each row of image plots represents different values of the rate of loss owing to mutation or recombinational deletion (α) while holding (γ) constant, while each column represents different values of the rate of incorporation (γ) while holding (α) constant. The highlighted image plot in the lower left-hand corner identifies the range of parameter values we believe are most realistic—both α and γ are close to zero.

zero, the longer the epidemic lasts and the longer simulations need to be before the asymptotic equilibrium is achieved.

All simulations were conducted in R v. 2.15.1 [44].

4. Results
We start by focusing on a slow, approximately human life history, simply because more is known about human genomes than for any other species. We set δS = 0.95 and τ = 1.016–0.95 = 0.066. After considering a slow life history, we explore the dynamics for life histories of other speeds.

(a) SI/E model for a human life history
The dynamics of the SI/E model are straightforward to understand, depending on only four parameters (equation (3.1)). The proportion of the population in the infected class decreases as the mortality rate, τ, increases. This is because the second term on the right-hand side of (3.1) increases in size as all newborns enter the susceptible class. As the infection rate, β, increases in relation to δS(δS − 1), then the proportion of infected individuals increases as the second term on the right-hand side of (3.1) becomes smaller. When the virus imposes no mortality cost, then the proportion of infected simplifies to 1 − τ/β for all values of τ < β.

(b) SI/E model for a human life history
Figure 2 provides complete results for the SI/E model across the full range of parameter values for α, γ, βE, and βC. We do not perturb δS, δE, δC or τ, as at this stage we focus on a species with a slow life history and we assume both the endogenous and exogenous retroviruses impose a mortality cost.

The frequency of the population in the I,E compartment increases as the rate of incorporation (γ) increases relative to the rate of loss (α), with the highest frequencies achieved when γ is high and α is low, a pattern observed in data (please refer to our reasoning for values of α and γ in §3d). In addition to the effects of α and γ, the rates of infection—βE and βC—affect the rate of incorporation (please refer to our reasoning for values of α and γ in §3d). In addition to the effects of α and γ, the rates of infection—βE and βC—affect the frequency of the population in the I,E compartment. When βE is small, few individuals enter the I,E compartment, even fewer making it through to the I,E compartment.

The most likely rates of the parameter space are highlighted in the bottom left-hand corner of figure 2, where rates of loss owing to mutation or recombinational deletion are close to zero and rates of incorporation are also small. This suggests that a large proportion of the population ends up in the I,E compartment under likely values of α and γ.

(c) SI/E/P,RNR,P,LTR model for a human life history
This model differs from the previous model by containing two recovered classes—one for individuals who gained immunity (EDI) to the exogenous virus from an endogenous retroviral insertion in the genome (which has subsequently undergone recombinational deletion, after the threat of the exogenous virus has passed, leaving the characteristic signature LTR) (RNR), and one for those that mounted a successful immune response (cellular/humoral immunity) to the exogenous retrovirus without endogenization (RNR). The two additional rates included in this model are θ and θ’ which determine the rate at which immunity is acquired to the exogenous virus. These rates determine the rate out of the I,E and I,E compartments, respectively. For individuals entering the RNR, immunity is not inherited, and their offspring return to the susceptible compartment. The RNR compartment is an absorbing state—there is no way out. All parameters and definitions used are summarized in table 1.

### Table 1. Parameters and definitions in the SIR model.

<table>
<thead>
<tr>
<th>parameter</th>
<th>definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>δS</td>
<td>survival rate per time step of an individual uninfected with a virus</td>
</tr>
<tr>
<td>1−δS</td>
<td>the proportional reduction in δS caused by the exogenous retrovirus</td>
</tr>
<tr>
<td>1−δE</td>
<td>the proportional reduction in δE caused by the endogenous retrovirus</td>
</tr>
<tr>
<td>τ</td>
<td>the birth rate per time step</td>
</tr>
<tr>
<td>βE</td>
<td>the infection rate of an individual with the exogenous virus</td>
</tr>
<tr>
<td>βC</td>
<td>the infection rate of an individual with the endogenous virus</td>
</tr>
<tr>
<td>γ</td>
<td>rate of incorporation</td>
</tr>
<tr>
<td>α</td>
<td>rate of loss of the incorporated virus</td>
</tr>
<tr>
<td>⃗θ</td>
<td>the rate at which immunity is acquired to the exogenous retrovirus</td>
</tr>
<tr>
<td>⃗θ’</td>
<td>the rate at which immunity is acquired to the endogenous retrovirus</td>
</tr>
</tbody>
</table>
The dynamics of the full model can be understood by building on insights gained from analysis of the $SI_X I_E$ model (figure 2). A typical run of the model, where a high proportion of the population ends up in the $R_{LTR}$ compartment, is displayed in figure 3. The exogenous and endogenous infections initially lead to a decrease in population size. This is because the population consists of a large proportion of infected individuals that have lower fitness than individuals in the susceptible or recovered compartments. As individuals transition into the $R_{LTR}$ recovered state the population starts to grow again and the proportion of individuals in the $R_{LTR}$ compartment converges to approximately 80%. When the $SI_X I_E$ model predicts a large proportion of the population is in the $I_E$ compartment at equilibrium, then a larger proportion of individuals will be expected to transition into the $R_{LTR}$ class for a given value of $u_0$ than when it predicts a smaller proportion. Similarly, a greater proportion of the population will end up in the $R_N$ compartment and subsequently back in the $S$ compartment for a given value of $\alpha \theta$ than when it predicts a smaller proportion. Similarly, a greater proportion of the population will end up in the $R_{LTR}$ compartment and subsequently back in the $S$ compartment for a given value of $\theta$ when the $SI_X I_E$ model predicts a large proportion of individuals in the $I_X$ class. Consequently, the dynamics of the full model are determined by the dynamics of the $SI_X I_E$ model as well as the relative values of $\theta$ and $\alpha \theta$. For a human life-style history, we found that a large proportion of individuals ended up in the $R_{LTR}$ compartment, regardless of the values of the transition rates (figure 4 for a value of $\phi_0 = 0.95$).

(d) $SI_X I_E R_N R_{LTR}$ model for faster life histories

The model describes the flow of individuals between compartments. If a large proportion of the population is to end up in the $R_{LTR}$ compartment, a greater proportion of individuals need to flow from left to right in figure 1c than vice versa. The primary process generating flow from right to left is reproduction, with newborn individuals from the $I_X$ compartment entering the susceptible compartment, along with newborns from the $I_E$ compartment that have lost the active endogenous virus at the locus in question due to recombinational deletion or mutation, and newborn individuals from the $R_N$ compartment. As we increase the speed of the life history, with reproduction ($t$) being greater than survival ($f_S$), we would expect to increase the flow of individuals from right to left. Figure 4 illustrates the effects of varying the parameter values of the $SI_X I_E R_N R_{LTR}$ model outlined in figure 1c. For fast life histories (where the birth rate ($\tau$) is high and survival ($f_S$) is low), the flow of individuals into the $S$ compartment gets sufficiently high that the growth rate of the susceptible
The proportion of individuals that enter the $R_{\text{LTR}}$ compartment displays signatures of retroviral infection [3]. Regardless of parameter values for all parameters (different styled lines), the proportion of individuals that enter the $R_{\text{LTR}}$ compartment is zero for fast life histories (speed of life history on the x-axis is $\phi_b$). Initial parameter values (solid lines): (a) $\gamma = 0.001$, (b) $\beta_x = 0.5$, (c) $\beta_e = 0.4$, (d) $\theta = 0.05$, (e) $\theta' = 0.01$, (f) $\alpha = 0.001$. The other lines represent repeated increases in each parameter by 10%.

(e) A modified $SI_{\text{LTR}}R_{\text{LTR}}$ model

Both $SI_{\text{LTR}}R_{\text{LTR}}$ models predict a large proportion of the population in the $R_{\text{LTR}}$ compartment for approximately human life histories (compare figures 4 and 6 for large values of $\phi_b$). For slow life histories, the modified $SI_{\text{LTR}}R_{\text{LTR}}$ produces dynamics that are very similar to those observed in the original $SI_{\text{LTR}}R_{\text{LTR}}$ model displayed in figure 3. However, the dynamics of the two models differ for faster life histories—compare for example figures 4 and 6 for values of $\phi_b < 0.9$. In the modified $SI_{\text{LTR}}R_{\text{LTR}}$ model (figure 3d), the only route back to the susceptible class occurs when $\theta$ is high, i.e. it is easy to acquire immunity to the exogenous retrovirus. This is illustrated in figure 5. For values of $\theta$ above 0.1, few individuals enter the $R_{\text{LTR}}$ class, suggesting that if immunity to the exogenous virus can be acquired easily, we would not expect to see incorporation (assuming incorporation provides a route to immunity).

When $\theta$ is low (i.e. it is difficult to acquire immunity to the exogenous retrovirus), all individuals end up in the $R_{\text{LTR}}$ class across a wide range of parameter values, even when $\alpha$ (rate of loss) and $\gamma$ (rate of insertion) are small and equal. This result is unaffected by the life history of the host, except close to the limit (when $\phi_b$ is close to zero or close to unity). Figure 6 shows how parameter values have relatively little influence on the proportion of the population that ends up in the $R_{\text{LTR}}$ compartment across the range of life histories.

5. Discussion

ERVs and their signatures represent a significant proportion of vertebrate genomes, yet the role of disease in driving endogenous retroviral signatures to high frequencies has not previously been explored. Through the novel application of epidemiological models, we show that endogenous retroviral signatures can reach high frequencies at a particular locus, if integration provides an immunological advantage and the rate at which immunity is acquired to the exogenous virus is much higher in individuals that have the incorporated retrovirus, than the rate at which immunity arises to the exogenous virus in those individuals without an incorporated virus.
Numerical analysis of our model suggests that an LTR at a locus can reach high frequencies when the rate at which immunity arises to the exogenous retrovirus is lower than the rate at which immunity arises with the ERV. This result provides theoretical support for the idea that endogenization can help a host evolve immunity to a retroviral infection [12,13,27,45]. There are a number of examples of ERVs conferring some immunity to related exogenous retroviruses through a variety of mechanisms. It has been demonstrated that in mice, expression of the Env protein from an ERV binds to the same cell surface receptors as a related exogenous retrovirus, effectively blocking the entry to the cell of the exogenous form [29]. Similar examples also exist in chicken and sheep [9,13,30]. Other examples target virus replication such as the Fv1 locus in mice [46,47]. A recent review highlights more examples of co-option of ERVs as a host defence mechanism to retroviruses [26]; with the abundance of genomic data available, there is no doubt that more examples will be discovered.

Our model reveals the range of parameter values that allow an LTR at a single location in the genome to reach fixation within a host population. But why do we see so many LTRs in many vertebrate genomes? We propose a hypothesis that once the endogenous process starts, an exogenous retrovirus incorporates at numerous locations in the host genome. At the majority (potentially all) of these locations the insertion provides no host benefit via promoters in the LTR or of the co-option of a viral gene. The ERV continues to produce virions until immunity (EDI) occurs or the exogenous counterpart is no longer a threat. At this point, mutation and recombinational deletion would build up at the loci where the virus had incorporated, leaving the LTR signatures of endogenous integration as the active virus is removed. If this process were to occur, then we would expect to see solo LTR formation occurring many times at the same locus, which has been previously described [8]. Additionally, if incorporation of new retroviral family was rare, then we might expect to see signatures of bursts of retroviral endogenization that occur during a period, followed by periods of inactivity. Existing data appear consistent with such a hypothesis. For example, the most recently active retroviruses in the human genome belong to the HERV-K family. The oldest members of this family integrated before the divergence of Old World monkeys and apes (approx. 25 Ma), but HERV-K elements have undergone several periods of expansion throughout primate evolution, with a number of insertions unique to humans, some of which are polymorphic [8,48–51]. However, there is little evidence of current active ERVs in screened individuals. It is possible that immunity to HERV-K has now evolved and spread through the population, that mutation and recombinational deletion have removed functional ERVs and we see numerous LTR signatures of past infections.

It is likely that there are other forces which would influence the subsequent spread through a genome after the initial invasion event by EDIs, for which our model does not account. For example, there may be interference between multiple insertions in the genome, and it is unlikely to be a straightforward additive effect of immunity: more insertions may not necessarily mean more immunity. As the various mechanisms of EDIs are better understood, these factors can then be incorporated into the model.

Most ERV insertions are ancient, and consequently it is difficult to say why those insertions were so successful at invading the host genome. In the absence of exogenous viral counterparts to these older insertions, we can only speculate as to how many of these ERV insertions would fall into the category of EDIs. There are very few real-time genome invasions, with an exogenous viral counterpart, to test this hypothesis. In humans at least (for which most data exist), for most HERV lineages, the main method of proliferation through a genome appears to be by recombination [52], which would be in line with what we propose. However, we acknowledge that there are also additional copying mechanisms that contribute to the spread of a few HERV families within the genome, such as retrotransposition [53], and then there are those ERVs that possess a degraded env gene and whose main method of proliferation now appears to be transposition [54], on which our model would clearly have no bearing.

It could be the case that the majority of ERVs present in genomes are neutral and have drifted to fixation, and there are a number of models that describe these cases [55–57]. Our results do not contradict genetic drift as the main evolutionary mechanism in driving ERV insertions to fixation: we merely provide an alternative for consideration. In those instances where the insertion does confer immunity (EDI), our results suggest that retroviral signatures can reach high frequencies if immunity cannot be achieved in any other way (cellular or humoral). Realistically, if there is even a tiny selective advantage that increases the frequency of these insertions in the population, then it would greatly reduce the chances of these insertions being lost to drift, and in reality it could be the case that it is a combination of some immunological advantage and genetic drift that leads us to observe the numerous signatures of past retroviral infections for those particular ERV lineages involved in EDI.

All models are wrong; they are designed to be simplifying caricatures of reality and make assumptions. However, they can be useful in posing hypotheses to explain patterns in data. One process we do not include is that upon initial invasion
of a genome by a retrovirus: the retrovirus would be present in the heterozygous state. We do not account for differences between heterozygotes and homozygotes as we are looking at conditions for the invasion of the insertion, and our results reveal that a large proportion of the population ends up in the $R_{\text{LTR}}$ compartment regardless of the value of $\alpha$, the parameter that will change as a function of the proportion of the population in the heterozygous/homozygous state.

Our model suggests potentially fruitful areas of future research. First, relatively little is known about rates of endogenization, of rates of mutation and recombinational deletion, and of rates of the acquisition of immunity outside of humans. Genomic studies on species of organism where ERVs are still active should help in answering these questions [9,10,13], as will comparative cross-species analyses of vertebrate genomes.

There are various other modifications we could make to the model that we do not explore here. Some viral infections impact fertility rates rather than survival rates [58–61]. We could incorporate such infections into our models by including compartment-specific fertility rates, much in the same way as we currently have compartment-specific survival rates. We strongly suspect that if we were to do this our general conclusions would not change unless the choice of compartment-specific fertility rates substantially impacted the rate of flow of individuals from right to left in figure 1d. The models we have developed do not incorporate specific mechanisms. We hope to incorporate more detailed immunological mechanisms into our future models, but in order to do that a better understanding is required of how immunity to retroviruses evolves, and the various mechanisms via which endogenization can aid the evolution of immunity, which may also vary with life history.

Under the scenario of retroviral insertions providing immunity (EDI), our model reveals the conditions under which a provirus involved in such immunity would spread to fixation. Once the threat of the exogenous retroviral counterpart is removed, the selection pressure upon retaining the provirus is removed, and these insertions are then free to undergo recombinational deletion, leaving the LTR signatures of endogenous retroviral infection that are so frequently observed in host populations. Our findings suggest that ‘neutral’ insertions are likely to reach high frequencies, or fixation, when hosts can more easily evolve immunity with an endogenous version of the retrovirus to the exogenous type. If future research provides support for this hypothesis, it is possible that retroviruses have played a major role in the evolution of the vertebrate immune system with regard to this form of immunity.

**Acknowledgements.** We would like to thank anonymous reviewers, Louise Johnson, Aris Katzourakis and Richard Nichols for comments and suggestions.
transcripts are selectively packaged into retroviral particles produced by the human germ cell tumor line Tera-1 and originate mainly from a provirus on chromosome 22q11.21. J. Virol. 82, 10 008 – 10 016. (doi:10.1128/JVI.01016-08)


41. Kong A et al. 2012 Rate of de novo mutations and the importance of father’s age to disease risk. Nature 488, 471 – 475. (doi:10.1038/nature11396)


