Integrating Genealogy and Epidemiology: The Ancestral Infection and Selection Graph as a Model for Reconstructing Host Virus Histories

David Welch\textsuperscript{a}, Wiremu Solomon\textsuperscript{c}, Allen Rodrigo\textsuperscript{b}, Geoff Nicholls\textsuperscript{a,*}

\textsuperscript{a}Department of Mathematics, University of Auckland, New Zealand
\textsuperscript{b}School of Biological Sciences, Allan Wilson Centre for Molecular Ecology and Evolution and Bioinformatics Institute, University of Auckland, New Zealand
\textsuperscript{c}Department of Statistics, University of Auckland, New Zealand

Abstract

We model the genealogies of coupled haploid host-virus populations. Hosts reproduce and replace other hosts as in the Moran model. The virus can be transmitted between individuals of the same generation and from parent to child. The epidemic model allows a selective advantage for susceptible over infected hosts. The coupled host-virus ancestry of a sample of hosts and their viruses is embedded in a branching and coalescing structure that we call the Ancestral Infection and Selection Graph, a direct analogue to the Ancestral Selection Graph of (Krone and Neuhauser, 1997a). We prove this and discuss various special cases. We use simulations to explore the level of infection within the population before it reaches equilibrium, where the infection has died out.
1 Introduction

In this paper we set up and analyse a model of the coupled host-virus genealogies of a panmictic host population. The model is an extension of the Moran model, an explicit model of mutation and selection. The genealogy process determined by the Moran model is well approximated by the ancestral selection graph process of Krone and Neuhauser (1997a). In a similar sense, the ancestral infection and selection graph process specified here approximates the genealogy and infection processes in our idealised host-virus populations.

We consider a population of \( N \) host individuals. Hosts are susceptible to, or infected with, a virus. Label susceptible individuals type \( A_S \) and infected individuals type \( A_I \). Infected individuals reproduce at constant rate \( \lambda_I \). Susceptible individuals reproduce at a possibly higher rate \( \lambda_S \). For some \( s_N \geq 0 \), let

\[
\lambda_S = \lambda_I(1 + s_N). \tag{1}
\]

Throughout this paper we treat fertility selection, that is, selective advantage arising from higher birth rate. Viral infection can lead to selective advantage based on lower death rate for susceptible individuals, so called “viability selection”. The models we present may be adapted to this case and lead to similar results.

* Corresponding author.

Email address: nicholls@math.auckland.ac.nz (Geoff Nicholls).
An uninfected parent cannot spontaneously produce infected offspring, so offspring of type $A_S$ parents are type $A_S$. However, an infected parent may vertically transmit the virus to its offspring. Suppose offspring of type $A_I$ parents are type $A_I$ with probability $(1 - u_N)$, and type $A_S$ otherwise. Non-transmission of infection from an infected parent to its child is one of a class of events which will correspond to a mutation event.

The virus is spread horizontally within the population via infectious contact events. These events are initiated by all hosts at constant rate $\lambda_1 c_N$, for some $c_N \geq 0$. The initiating host (the \emph{initiator}) chooses another host (the \emph{target}) uniformly at random from the host population, including itself. The virus may pass between the two depending on their types immediately prior to the event. If just one of the two individuals is infected, both individuals emerge infected. If either both are infected, or both are susceptible, the contact event has no effect. See Table 2. Regarding contacts between infected individuals as ineffectual is equivalent to assuming no super-infection of hosts. In section 4 we discuss an asymmetric scheme where only infected hosts may initiate contact.

<table>
<thead>
<tr>
<th>Type before contact</th>
<th>Type after contact</th>
</tr>
</thead>
</table>
| \begin{tabular}{llll}
Initiator & Target & Initiator & Target \\
S & S & S & S \\
S & I & I & I \\
I & S & I & I \\
I & I & I & I \\
\end{tabular} |

(2)

Finally, host individuals arrive via a migration process at rate $\lambda_I \beta$. Migrants
are susceptible with probability \( p \) and otherwise infected. Migrants replace an individual chosen uniformly at random from the population of hosts. We are making the simplifying assumption that the population from which the migrants are drawn is not itself subject to the same stochastic evolution as the population they enter.

We represent the level of infection in the population at time \( t \) as a continuous time Markov chain \( Y(t) = (Y_S(t), Y_I(t)) \), where \( Y_k(t) \) is the number of individuals of type \( A_k \) at time \( t \), \( k = S, I \). Clearly, \( Y_S(t) + Y_I(t) = N \). If \( Y_S(t) = j \), \( j \in \{0, \ldots, N\} \), the following transitions occur:

\[
j \rightarrow \begin{cases} 
  j + 1 & \text{at rate } \lambda_{Sj} \frac{(N-j)}{N} + \lambda_{Ij} u_N (N-j) \frac{(N-j)}{N} + \lambda_I \beta p \frac{(N-j)}{N} \\
  j - 1 & \text{at rate } \lambda_I (1-u_N) (N-j) \frac{j}{N} + 2 \lambda_J c_N j \frac{(N-j)}{N} + \lambda_I \beta (1-p) \frac{j}{N}
\end{cases}
\]

In order to simplify the study of the birth-death process (3), we analyse instead the diffusion approximation of the related scaled proportion processes, \((Y_I(t)/N, Y_S(t)/N)\) in the limit \( N \to \infty \) (see, for example, Ewens (1979)). Following Krone and Neuhauser (1997a), units of time are chosen so that \( \lambda_I = N/2 \), and we suppose there exist constants \( C < \infty \) and \( \gamma > 0 \) and real scalars \( \mu, \sigma \) and \( \theta \) so that

\[
|Nc_N - \mu| \leq CN^{-\gamma} \\
|Ns_N - \sigma| \leq CN^{-\gamma} \\
|Nu_N - \theta| \leq CN^{-\gamma}
\]

for all sufficiently large \( N \).

In the limit \( N \to \infty \), the proportion of susceptible individuals, \( Y_S(t)/N \), is a
diffusion process, $W(t)$ on $[0, 1]$, with drift

$$a(x) = ((\theta + p\beta)(1 - x) - (1 - p)\beta x - 2\mu x(1 - x) + \sigma x(1 - x))/2 \quad (5)$$

and diffusion $b(x) = x(1 - x)$. When $(1 - p)\beta > 0$ and $\theta + p\beta > 0$, the boundaries at $x = 0$ and $x = 1$ are reflecting. In that case the equilibrium density $h(x), x \in [0, 1]$ for a realisation, $W(t) = x$, of the proportion process is given by Wright’s formula (see for example Ewens (1979))

$$h(x) = Kx^{\theta+p\beta-1}(1 - x)^{(1-p)\beta-1} e^{-(2\mu-\sigma)x}, \quad (6)$$

where $K$ is a normalising constant.

The process defined above is related to the Moran process. Consider a Moran process with susceptible and infected individuals reproducing and replacing one another uniformly at random over the population. In such a process, susceptible individuals generate infected individuals by an event in which a susceptible gives birth, mutates from susceptible to infected, and replaces a susceptible individual. This leads to term like $\lambda_S uN^2/N$ in the rate for $j \rightarrow j - 1$. In contrast, in the model above, susceptible individuals generate infected individuals by initiating contact events with infected individuals. As a consequence, (3) contains the term $\lambda_I cN j(N - j)/N$. Direct mutation from susceptible to infected is represented by infected host migrants replacing susceptible hosts, a process with rate $\lambda_I \beta pj/N$. The models have the same diffusive limit for the proportion of susceptibles, up to a reparameterization. If, in the scaled Moran process, $\theta_I$ is the mutation rate from $A_I$ to $A_S$, $\theta_S$ the rate for the reverse event, if $\hat{\sigma}$ is the selective advantage of type $A_S$ over $A_I$, and $X(t)$ is the diffusion on $[0, 1]$ of the proportion of type $A_S$ individuals in the population, then $X(t)$ has drift $a(x) = (\theta_S(1 - x) - \theta_I x - \hat{\sigma} x(1 - x))/2$.
and diffusion $b(x) = x(1 - x)$. The diffusions $W(t)$ and $X(t)$ are equivalent when $\theta_S = \theta + p\beta$, $\theta_I = (1 - p)\beta$ and $\sigma = 2\mu - \sigma \geq 0$. Differences between the genealogy-graph processes determined by the two models are discussed in Section 2.

The idealised infection is never truly endemic. In the absence of immigrant infection, when $p = 1$ or $\beta = 0$, the right hand boundary, $W(t) = 1$, is attainable and absorbing, in the sense of Karlin and Taylor (1981), and therefore an exit boundary. Regardless of the initial state, it is reached in finite expected time with probability 1. Since the drift and diffusion terms are zero there, the process remains in that state after reaching it. In section 4, we explore the process using numerical simulations. In these simulations, and others which we do not report, we find that, for values of the the parameters $\mu$, $\sigma$ and $\theta$ which are at least plausible, the infection persists for several times $N$ generations without immigration. This time will often be large compared to the time scale over which the background parameters of the biological population can be assumed constant. In the absence of immigrant infection, the presence of infected host in a sample indicates the process is not in equilibrium. We will see that this is not in itself an obstacle to statistical inference (i.e., parameter estimation). The role of immigration in the model is not therefore to impose a spurious persistence for the disease, but rather to allow host and viral ancestral lineages to terminate outside the model-population. This is discussed further in Section 2.
2 Graphical Representations

A realization of the infection process (3), acting in a population of hosts, can be represented via a percolation diagram. A realization of the history of the infection for a sample of hosts is a subgraph of this diagram. In the limit (4) of large populations, the subgraph process converges to a graph process we define below. We call this limiting graph process the ancestral selection and infection graph process. The proof can be adapted almost unchanged from Krone and Neuhauser (1997a). These authors set up a percolation-diagram representation for the Moran model, and established the corresponding limiting ancestral selection-graph process.

2.1 The Forward Model

A realization of the infection process is simulated as follows. Refer to Fig. 1 for an instance. Let $I = \{1 \ldots N\}$ be the set of $N$ site labels for a population of $N$ hosts. The sites correspond to the vertical lines in Fig. 1 with time increasing down the page. Arrows representing birth and contact events are drawn between the sites on the space $I \times [0, \infty)$. The times and types of these events are simulated by thinning independent Poisson processes. This is done in such a way that the individual occupying any particular site encounters events at a rate appropriate for their infection type.

For each ordered pair of sites $(i, j)$ we simulate a Poisson process with rate $\frac{\lambda_f(1+s_N+c_N)}{N}$. At each arrival time $t$ we draw an arrow from site $i$ to site $j$ at time $t$. To decide the type of this event we draw a uniformly distributed random variable, $v \in [0, 1]$. If $v < \frac{\lambda_f}{\lambda_f(1+s_N+c_N)}$ label the arrow $I$, if $v \in$
\([\lambda_I(1+s_N+c_N), \lambda_I(1+s_N+c_N)]\) label the arrow \(S\), otherwise \(v \in [\lambda_I(1+s_N+c_N), 1]\) and we make the arrow double-headed and label it \(C\).

At an \(I\)-arrow from site \(i\) to site \(j\) offspring from a birth at site \(i\) replaces the individual at \(j\). This event occurs irrespective of the type at \(i\). At an \(S\)-arrow from site \(i\) to site \(j\), offspring from a birth at site \(i\) replaces the individual at \(j\) but only if the type of the individual at site \(i\) is \(A_S\). This ensures that type \(A_I\) individuals encounter birth events at rate \(\lambda_I\) and type \(A_S\) at rate \(\lambda_I + s_N \lambda_I\).

Each time we simulate an \(I\) arrow (from \(i\) to \(j\) say) we simulate an independent event with probability \(u_N\) and place a dot on the arrow if this event occurs. If the individual at \(i\) is type \(A_I\) and the dot is present, the offspring of \(i\) replacing the individual at \(j\) is type \(A_S\). In all other cases the offspring is the same type as the parent. In this way we simulate non-transmission of infection at birth.

For each site \(i = 1, 2 \ldots N\) we simulate a migration process at rate \(\lambda_I \beta / N\). Each arrival of the migration process is indicated by a dot, and marked \(S\) with probability \(p\) (the host individual at site \(i\) is replaced by a type-\(A_S\) host) and otherwise \(I\) (replacement by a type-\(A_I\) host). We will refer to the dot process (non-transmission events on \(I\)-arrows and \(I\)- and \(S\)-immigration events) as the mutation process.

A \(C\)-arrow connecting sites \(i\) and \(j\) represents contact between the individuals at \(i\) and \(j\). The infection may pass between them according to the rules in Table 2. We make the arrow double-headed to indicate that the virus may travel in either direction along the arrow.

If at a given time we know the type at every site, we can propagate the types forward in time, using the percolation diagram to keep track of both the host
and viral genealogies.

2.2 The Ancestral Infection and Selection Process

Where data is available from a small sample $n \leq N$ of individuals from a much larger population, it is convenient to simulate ancestral and infective history for subsets of individuals without simulating the corresponding history for the entire population. Following Krone and Neuhauser (1997a), we define a dual percolation process which gives us this information. A realization of the dual percolation process is a subgraph of the percolation diagram realized by the forward process. The example pictured in Fig. 2 is derived from the realization of the forward process depicted in Fig. 1.

We obtain a realization of the dual percolation process as follows. Using the forward process, simulate a percolation diagram from time $t = 0$ to time $t = T$ omitting mutation events. Define a new time scale, “dual time”, which increases into the past with dual time equal to $t_0$ at time $T$. Reverse all arrows in the forward process so that arrows point to ancestors. Consider a sample of $n$ hosts drawn from the $N$ at dual time $t_0$. Beginning at $t_0$ at the sites corresponding to the $n$ individuals in the sample, trace the ancestral lineages of the sample hosts and their infections back through the percolation diagram to dual time $t_0 + T$. Traced lineages branch and coalesce and thereby determine a subgraph of the full percolation diagram. This subgraph is a realization of the dual percolation process. We illustrate the process for the $n = 3$ individuals at sites 2, 5 and 8 at dual time $t_0$ in Fig. 2.

Consider site 2 in Fig. 2. At dual time $t_1$ the offspring of the individual at site
Fig. 1. A percolation diagram realization of the infection model for $N = 8$. If at time $t = 0$, individuals at sites $\{1, 3, 5, 7\}$ are susceptible and the rest infected then, propagating types according to the rules set out in the text, at $t = T$ individuals at sites $\{4, 5, 6\}$ are susceptible and the rest infected. There are no immigration events in this example.

Fig. 2. The dual process for the infection model obtained by reversing the direction of time and the direction of the arrows in Fig. 1.
1 replaces the individual at site 2. This is an \( I \)-birth, so it occurs irrespective of the type of the individual at site 1. In the dual process, the individual at 1 is the ancestor of the individual we are tracing, so we follow the arrow from site 2 to 1. Similarly, we follow the \( I \)-arrow at \( t_4 \) to an ancestor at site 3. We ignore the arrow at \( t_6 \). It is incoming in the dual, corresponding to a birth from site 3 in the forward process. At \( t_8 \), a \( C \)-arrow connects site 3 and site 7. Infection could have entered site 3 at this time depending on the types of the two individuals in contact. We do not know the types but we wish to keep track of the ancestry of any infection, so we branch at this point. We follow the potential infection-ancestry to site 7 while at the same time following site 3. Continuing back in time, we follow the \( I \)-arrows at times \( t_9 \), \( t_{10} \) and \( t_{11} \). We approach time \( t_{12} \) following paths at sites 3 and 6. At time \( t_{12} \), we find that the ancestor of the lineage at site 6 came from site 3, so we follow the arrow from site 6 to site 3. The two lineages coalesce to a single lineage. The history of the individual at site 5 at time \( t_0 \) is retraced in a similar way. At time \( t_2 \), we encounter an \( S \)-arrow in the dual. The ancestor of the individual at site 5 could come from site 6 if site 6 was type \( A_S \) at this time. Since we do not know the types of the individuals we follow both paths, 5 and 6. Continuing in this way, we obtain the infective history shown in bold in Fig. 2.

Because the forward process is reversible, and events on each site are realized independently, the dual subgraph may be simulated in dual time from the \( n \) sites present at \( t_0 \) to arbitrarily large dual times. There is no need to simulate events in lineages outside those in the dual subgraph or work on a percolation diagram realized by the forward process with \( T \) fixed. For the applications we have in mind, it is convenient to stop the dual process at the first (coalescence) time the number of lineages in the dual process becomes one. In Fig. 2 that
time is $t_{12}$. The individual at the point of coalescence is ancestral to all the sample individuals. Any infection it carries is ancestral to any infection they carry. Let $t_{UA}^N$ denote the dual time of this joint ultimate ancestor.

Note that immigration events terminate host and viral ancestry in dual time. In our specification of the dual percolation process we continue tracing these lineages. This is unnecessary but harmless. It leads to a standard graphical representation with a single connected graph and a single ultimate ancestor.

Mutation events may be simulated on the dual subgraph. Given the infection type of the joint ultimate ancestor, infection type is then determined at all points on the dual subgraph. For example, in Fig. 2, if the individual at site 3 is infected at dual time $t_{12}$ then at dual time $t_0$, the individuals at sites 2 and 8 are infected, while the individual at site 5 is susceptible. Once infection types have been propagated down the dual subgraph, ancestry of hosts at $S$-arrows and ancestry of infection at $C$ arrows is decided. The genealogy of the individuals in the dual subgraph at dual time $t_0$, and the genealogy of any infection they carry, can then be traced back to dual time $t_0 + T$.

Having described the dual percolation subgraph process we now define a near equivalent graph process containing just those events in the subgraph process which are needed for the propagation of infection type down the graph, and ancestry up the graph. Site labels are dropped from the dual subgraph and paths made up of sequences of $I$-arrows are represented by a single edge. Coalescing and branching events, and branching event types ($C$ or $S$) are recorded along with the dual times for all events. We call this cut-down realization the dual graph. The dual graph at left in Fig. 3 summarizes events shown in the bold subgraph of Fig. 2.
Denote by $\mathcal{G}_{N,n}$ the process realizing dual graphs for samples of size $n$ drawn from a population of size $N$. The graph on the left in Fig. 3 is a realization of $\mathcal{G}_{8,3}$. Note that a realization of $\mathcal{G}_{N,n}$ does not include mutation events, or details of infection type or ancestry. The mutation process on the dual percolation subgraph determines a mutation process $\mathcal{Y}_{N,n}$ on realizations of $\mathcal{G}_{N,n}$, i.e., on dual graphs.

Rules for propagating infection type and ancestry through branching and mutation events in the dual graph are given in Figures 4 and 5.

Consider a branching event at dual time $t$. Immediately below the branching event, at dual time $t^-$, we have a single edge, which we refer to as the branching edge. Immediately above the branching event, at dual time $t^+$, we have two edges. One of these two edges is labelled the continuing edge (it corresponds to the path in the dual subgraph that continues on the original site) and the other incoming. We place an arrow on the incoming edge where it connects to the branching edge. At an $S$-branch, the host on the edge immediately below the branching, at $t^-$, is descended from the host on the incoming edge if the incoming edge is type $A_S$ at $t^+$. At a $C$-branch, infection immediately below the branching, at $t^+$, is descended from infection on the incoming edge if, at $t^+$, the incoming edge is type $A_I$ and the continuing edge is type $A_S$. Black dots on the dual graph are either unlabelled and represent non-transmission of the virus or are labelled and represent an immigration event. The type on an edge immediately below an unlabelled black dot is therefore $A_S$, regardless of the type above the dot. The type below a labelled dot is type $A_S$ or $A_I$ depending on whether the label is $S$ or $I$. The ancestry of both the infection and the host below a labelled black dot is unrelated to the ancestry above the dot. In Fig. 3, simulated dual-graph genealogies for sampled host and sampled infection
determined by the percolation subgraph shown in Fig. 2 with an ultimate ancestor of type $A_I$ are shown.

The dual-graph and mutation processes, $\mathcal{G}_{N,n}$ and $\mathcal{Y}_{N,n}$, together converge in distribution, in the limit $N \rightarrow \infty$ defined in (4), to processes $\mathcal{G}_n$ and $\mathcal{Y}_n$. We refer to $\mathcal{G}_n$ as the ancestral infection and selection graph process (AISG-process) and to $\mathcal{Y}_n$ as the mutation process. Instantaneous rates for events in $\mathcal{G}_n$ are given in terms of the parameters of the diffusion process $W(t)$, defined below (4), as follows. If at dual time $t$ the AISG-process has $k$ lineages, then each pair of lineages coalesces at rate 1, each lineage $C$-branches at rate $\mu$ and $S$-branches at rate $\sigma/2$. A realization of $\mathcal{G}_N$ starts at time $t_0$ with $n$ lineages and terminates at the first time that $k$ becomes one, $t_{UA}$ say. The $\mathcal{Y}_n$-process realizes non-transmission, and $S$ and $I$ immigration events, independently on each lineage at rates $\theta/2$, $p\beta$ and $(1-p)\beta$ respectively.

Given the infection type of the joint ultimate ancestor, the infection type and ancestry at all points on all lineages of a realization of $\mathcal{G}_n$ and $\mathcal{Y}_n$ is determined according to the rules set out in Fig. 5. A detailed justification that $\mathcal{G}_n$ and $\mathcal{Y}_n$ are the limiting processes of the dual graph and mutation processes follows Krone and Neuhauser (1997a) closely, and is therefore omitted. The following summary emphasizes points of difference.

A realization of $\mathcal{G}_{N,n}$ may contain events of a type which cannot be generated by $\mathcal{G}_n$. These are vertices of degree four, corresponding to events in the dual process in which an $S$ or $C$ arrow links two lineages already in the dual. These events are called collisions. We stipulate that at each collision, a single fictitious lineage is created and is allowed to evolve like all other lineages. If, at dual time $t$, graph process $\mathcal{G}_{N,n}$ has $k$ lineages, each lineage encounters $S$
collisions at rate $ks_N/2$ C collisions at rates $kc_N$ and $S$ and $C$ branchings at rates $(N - k)s_N/2$ and $(N - k)c_N$. Each pair of lineages in $G_{N,n}$ coalesce at rate 1.

Consider a dual graph process, $G_{N,n}^*$, derived from a Moran process (as described in Sec. 1) with selective advantage $s_N^* = 2c_N + s_N$. If $G_{N,n}^*$ has $k$ lineages, each lineage branches at rate $(N - k)s_N^*/2$ and encounters collisions at rate $ks_N^*/2$. At collisions, a single fictitious particle is created. Each pair of lineages coalesces at rate 1. If we ignore the labels on branches and collisions in $G_{N,n}$ then $G_{N,n}^*$ and $G_{N,n}$ are identically distributed. Krone and Neuhauser (1997a) show that in the limit $N \to \infty$, $G_{N,n}^*$ is the ancestral selection graph $G_n^*$ with branching rate $\sigma^*/2 = (2\mu + \sigma)/2$. The $S$- and $C$-branch (collision) labels can be independently imposed on $G_{N,n}$ and $G_n$, i.e., a branch (collision) is labelled $S$ with probability $\sigma/(2\mu + \sigma)$, otherwise, it is labelled $C$ with probability $2\mu/(2\mu + \sigma)$. It follows immediately that in the limit $N \to \infty$, $G_{N,n} = G_n$.

The dual mutation process $\mathcal{Y}_{N,n}$ may include non-transmission events at the same dual time as coalescing events in $G_{N,n}$. These simulataneous events are almost surely absent from the $\mathcal{Y}_n$ process. Krone and Neuhauser (1997a) show that the probability of these events occuring in $\mathcal{Y}_{N,n}$ tends to 0 as $N \to \infty$. The immigration events in $\mathcal{Y}_{N,n}$ and $\mathcal{Y}_n$ occur as independent Poisson process along the edges of $G_{N,n}$ and $G_n$. The process $\mathcal{Y}_{N,n}$ is the combination of these two "mutation" processes (i.e., non-transmission and immigration). It follows from the above remarks that $\mathcal{Y}_{N,n}$ and $\mathcal{Y}_n$ are equivalent in the limit $N \to \infty$.

Notice that the features by which the ancestral selection graph and the AISG differ, namely in the rules for the propagation of types down the tree and
ancestry up the tree, play no part in the discussion of the limit graph process.

The AISG process $\mathcal{G}_n$ resembles the ancestral influence graph (AIG) process of Donnelly and Kurtz (1999). The AIG process models the joint genealogies of two genes at linked loci observed in a sample of individuals from a population subject to selection and recombination. The ancestral selection graph process can be thought of as a special case of both AIG and AISG processes. As in the AISG, one AIG contains two intertwined genealogies. Also, branching events in the ASG process are of two types, accounting in the AIG for the effects of selection and recombination. However, lineages in the AIG are identified as ancestral at one or both loci. The instantaneous dynamics of the ASG and AISG differ in the following way: in the AIG-process recombination branchings occur only on lineages ancestral at both loci; in the AISG lineages are not distinguished, so that contact branchings occur at a constant rate on all lineages. For this reason we have not looked for any simple mapping from closed form results for expectations in the AIG-process and those of the AISG-process. The exception are those properties shared by both processes and the ASG itself.

3 Likelihood

The $\mathcal{G}_n$ and $\mathcal{Y}_n$ processes determine a likelihood for the branching and non-transmission parameters $\sigma$, $\mu$ and $\theta$ for infection-type data on the leaves. We write down the likelihood below.

Let $\Gamma$ denote the space of all realizations $g$ of the AISG process $\mathcal{G}_n$ on $n$ leaves. Let $V$ denote the set of node labels for $g$. Let $L$ and $T$ denote the subsets of $V$
made up all leaf and internal node labels, and let $S$, $C$ and $Y$ denote the subsets of $T$ made up of all selection, contact and coalescent node labels. If $|X|$ is the number of elements in set $X$ then $|L| = n$ and $|T| = 2|S| + 2|C| + n - 1$. Node labels are indexed in order by time from 1 to $|T|$ so that for any $u, v \in V$, $u > v$ implies $t_u \leq t_v$. For $v \in L$, let $D_v \in \{A_I, A_S\}$ denote the infection type datum at leaf $v$. Let $E = \{(u, v) : u, v \in V, u < v, \langle u, v \rangle \text{ is an edge of } g\}$ denote the edge-set of $g$. For edge $\langle u, v \rangle \in E$ and $t \in [t_u, t_v]$, let $A_{\langle u, v \rangle}(t) \in \{A_I, A_S\}$ denote the value of the infection type process at time $t$ on edge $\langle u, v \rangle$. Let $B_{\langle u, v \rangle} = (B_{\langle u, v \rangle, 1}, B_{\langle u, v \rangle, 2})$ record the types entering and leaving an edge, that is,

$$B_{\langle u, v \rangle} = (A_{\langle u, v \rangle}(t_+), A_{\langle u, v \rangle}(t_-)).$$

For $w \in V$, $t_w^+$ denotes the usual limits above and below at dual time $t_w$. Let $B = \{B_{\langle u, v \rangle}\}_{\langle u, v \rangle \in E}$. Possible type values $B$ are constrained to satisfy the rules of Fig. 4 and Fig. 5, and must agree with type data $D$ at leaf nodes. Let $\mathcal{B}_{D, g}$ denote the set of allowed type values.

It is necessary to make a change in the units of the time variable. The individual hosts associated with the leaves may have been sampled at different calendar times. The units of the dual time variable $t$ are determined by the choice $\lambda_I = N/2$: one unit of dual time equals the mean time for $N/2$ generations to pass, and $N$ is unknown. Let $\rho$ equal the mean calendar time for one host generation. In the limit Eqn. 4, this is independent of host infection type. Let $\Theta = \frac{N}{2\rho}$ so that $\tau = \Theta t$ is measured in calendar units.

Let $\pi_{UA} = (\pi_I, \pi_S)$ give the probability distribution for the type of the joint ultimate ancestor in the absence of data. Let $\pi_{eqm}$ denote the equilibrium of the type-proportion process $W(t)$ defined at (5). The case $(1 - p)\beta = 0$ is
special. There is no immigration of type $A_I$, the epidemic is always transient and $\pi_{eqm} = (0, 1)$. If the host sample data contains type $A_I$, then the process (3) was not in equilibrium at the time of sampling so $\pi_{UA} \neq \pi_{eqm}$. The joint ultimate ancestor is certainly type $A_I$, and we may impose $\pi_{UA} = (1, 0)$, since there is no other mechanism for type $A_I$ to enter the process. When $(1 - p)\beta > 0$ it follows from results in Slade (2000) that $\pi_{UA} = \pi_{eqm}$. In that case $\pi_S = \int_0^1 x h(x) dx$ where $h(x)$ is given in (6).

Let $Q$ be a $2 \times 2$ matrix of rates, with $Q_{1,2} = (1 - p)\beta/2$ giving the rate for transitions $A_S \rightarrow A_I$ on an edge (i.e., a host lineage) in $g$, and $Q_{2,1} = (p\beta + \theta)/2$ the rate for $A_I \rightarrow A_S$. Note that, as Slade (2000) makes clear, $\pi_{eqm}Q \neq 0$ in general. The type at the joint ultimate ancestor and the type at any single leaf are (correlated) draws from the equilibrium of $W(t)$ given by Wright’s formula (6). This equilibrium distribution is preserved down $g$ not by the mutation process $\mathcal{Y}_n$ alone, but by the combined action of mutation and branching processes.

Let $k_i$ denote the number of lineages present during $[\tau_i, \tau_{i+1})$, the $i$th interval. Let $2\Theta R_i = k_i(k_i - 1) + \sigma k_i + 2\mu k_i$ define the total rate $R_i$ for events in that interval.

The likelihood for $\Theta, \mu, \sigma$ and $Q$ (i.e., $p, \beta$ and $\theta$) for infection type-data $D$ is

$$P\{D|\Theta, \mu, \sigma, Q\} = \int \sum_{B \in B_D, g} P\{B|g, Q\} P\{g|\Theta, \mu, \sigma\} dg, \quad (7)$$

where $dg = \prod_{v \in T} d\tau_v$ with counting measure over distinct $g$-topologies. In (7)

$$P\{g|\Theta, \mu, \sigma\} = \left(\frac{1}{\Theta}\right)^{n-1} \left(\frac{\sigma}{2\Theta}\right)^{|S|} \left(\frac{\mu}{\Theta}\right)^{|C|} |T|^{-1} \prod_{i=1} P e^{-R_i(\tau_{i+1} - \tau_i)}$$
and
\[ P\{B|g, Q\} = \prod_{(i,j) \in E} \left[ \exp(Q(\tau - \tau_i)) \right]_{\partial(i,j), 2, \partial(i,j), 1}. \]

4 Discussion

The infection model and AISG-process above is just one of a class of models for the joint ancestry of host and parasite. In this section we introduce some modifications to the infection model and discuss the related modified AISG-processes. In particular, we look at a model where there is no immigration and one where the contact process is asymmetric. In the case of no immigration, we provide the results of numerical simulations estimating the length of time that an infection persists in a population for certain parameter values.

The immigration process given above has the infection entering via host immigration. This has the effect of terminating host ancestry at both infected and susceptible host immigration events. Immigration of infection only (at rate \( \lambda_I \beta^* \)) does not terminate host ancestry and leads to \( Q_{1,2} = \beta^*/2 \) and \( Q_{2,1} = \theta/2 \) in (7). This type of immigration model is useful when studying, for example, a virus that regularly crosses a species barrier and establishes itself in a new host.

When there is no immigration, we assume that the joint ultimate ancestor is type \( A_I \) and that any infection in the sample can be traced back to this source. It is important in this scenario that the infection remains within the population for a time comparable to the time to the joint ultimate ancestor, \( t_{U_A} \) so that the outcome, that is, the data, is a likely event. If the infection typically dies out in a time small compared to \( t_{U_A} \) this is a sign of some model
mispecification. If $\beta = 0$, the diffusion process $W(t)$ describing the proportion of susceptibles in the population has drift $a(x) = (\theta(1-x) - (2\mu-\sigma)x(1-x))/2$ which has zeros at $x = 1$ and $x = \frac{\theta}{2\mu-\sigma}$. Since $x \in [0,1]$, $a(x) > 0$ when $\frac{\theta}{2\mu-\sigma} < 0$ or $> 1$ and, in these cases, the process is attracted to the absorbing endpoint $x = 1$. The process remains on the interior of the space longest when $\frac{\theta}{2\mu-\sigma} \in (0,1)$ and $2\mu - \sigma$ is large. In this case, $\frac{\theta}{2\mu-\sigma}$ is a stable fixed point of the drift coefficient and the denominator $2\mu - \sigma$ determines the strength of the restoring drift to that point.

The expected time that the infection remains in the population is the expected hitting time of the absorbing boundary, $W(t) = 1$. In Fig. 6, we provide the results of a simulation study to investigate this value for two sets of parameter values. Fig. 6(a) shows results for parameter values $\mu = 1.5$, $\sigma = 0$, $\theta = 1.3$ and indicate that the expected hitting time of the boundary, starting at the midpoint $x = 0.5$, is approximately $4.5(2)$ time units (i.e., $4.5N$ generations). Fig. 6(b) shows results for values $\mu = 4$, $\sigma = 1$, $\theta = 2$ and put the expected hitting time of the boundary, starting from 0.5, at approximately $13.2(4)$ time units.

The expected value of the time to the joint ultimate ancestor, $t_{UA}$, can be calculated using the following expression derived by Krone and Neuhauser (1997a):

$$E_n[t_{UA}] = 2(1 - \frac{1}{n}) + 2 \sum_{r=1}^{n-1} \frac{1}{r(r+1)} \frac{e^{2\mu+\sigma}}{(2\mu+\sigma)^{r+1}} \int_0^\sigma t^{r+1}e^{-t}dt,$$

where $n$ is the sample size and $2\mu + \sigma$ is the total branching rate. The value of $E_n[t_{UA}]$ increases rapidly as the sum $2\mu + \sigma$ increases. For a sample of two individuals as in (a) with $2\mu + \sigma = 3$, $E_2[t_{UA}] \approx 3.75$, while when $2\mu + \sigma = 9$, as in (b) $E_2[t_{UA}] \approx 200$. In case (a) there is a reasonable chance the
sampled hosts are infected given infection at the ultimate ancestor. In case (b) the data would not typically contain infection inherited from the ultimate ancestor. Immigration is likely to play an important role in any real statistical inference when $2\mu + \sigma$ is large (as will often be the case). When $t_{UA}$ is large compared to the time scale determined by $\mu$ and $\sigma$, we should be able to estimate parameters $\mu$ and $\sigma$ without reconstructing the AISG back to $t_{UA}$. Population size variation could also play a role, since it will reduce $t_{UA}$.

The last case we consider is one where contact between hosts is no longer symmetric. This scheme is appropriate when modelling a virus which causes the host to act in a certain way that spreads the virus, for example, causing the host to sneeze. The contact scheme characterised by table 2 makes certain assumptions about the nature of contact between hosts that may not be suitable in all biological contexts. Namely, the assumptions that all hosts initiate contact and that the parasite may equally easily be passed from the initiator to the target as vice versa. It is easy to check that if only infected hosts can initiate contact, or equally, only contact initiated by infected hosts has any effect, the only change in the birth-death process (3) is that the second term in the rate of decrease is halved, becoming $\lambda_{CN}j(N - j)/N$. This rescaling carries throughout the analysis and the AISG remains unchanged except for the resulting lower C-branching rate, $\frac{nK}{2}$.

5 Conclusions

We have presented a model of the spread of a vertically and horizontally transmitted virus in a panmictic haploid host population of constant size. The model allows for a selective advantage of susceptible hosts over infected hosts.
The parameters that define the model are the population size, the transmission probability of the virus from parent to offspring, the rate of infectious contact by which the virus is spread horizontally, the rates of immigration of hosts and infection and the level of selective advantage held by the susceptible hosts.

We have shown how to construct the AISG which contains all information about the genealogy and the infection-genealogy of a sample of hosts, some of which may be infected. We noted that while the rules for the propagation of type and ancestry on the graph differ from the corresponding rules for the ancestral selection graph, the graph processes themselves are identical (up to a parameterization). Thus, some results such as the expected time to the ultimate ancestor derived for the ancestral selection graph may be used in the context of the AISG.

We provided an expression for the likelihood of a sample configuration given knowledge of the model parameters. Given the likelihood, it is a straightforward matter to derive an expression for the posterior distribution of the model parameters in a Bayesian framework. It seems likely, however, that for any accurate inference of these parameters and AISG topologies associated with a given sample, more data will be necessary than the basic susceptible/infected sample configuration. This data could include sequence data from both host and virus samples and any relevant information about the host ecology and viral epidemiology.
Acknowledgements

This research was partially supported by grants from the US National Institutes of Health, the Allan Wilson Centre for Molecular Ecology and Evolution, New Zealand and the University of Auckland Vice-Chancellor’s Development Fund. David Welch was supported by a University of Auckland Doctoral Scholarship.

References


Fig. 3. The dual graph at left is derived from the dual percolation subgraph in bold in Fig. 2. Mutation events on the dual percolation subgraph in Fig. 2 determine corresponding events on the dual graph at left in 3. The individual at the top of the graph is infected. Tracing infection type down the graph, we obtain the graph at centre left. Bold lines indicate infected lineages. Given infection type at the leaves we can trace (centre right, bold subtree) the genealogy of the virus and (right, bold subtree) host up the tree from the tips.
Fig. 4. Rules determining infection type and ancestry below a branching event at time $t$ in the dual graph: infection types at dual times $t^+$ and $t^-$ are indicated above and below the branching; an arrow indicates the incoming edge, the thick lines follow host ancestry, the dashed lines follow infection ancestry. Top row: rules for contact branchings. Bottom row: rules for selection branchings. In $C_4, S_1, S_3$ and $S_4$, no infection is present on the branching edge at $t^-$ so no infection ancestry is indicated.
Fig. 5. Rules determining infection type and ancestry below a mutation event (indicated by a dot). The thick lines follow host ancestry and the dashed lines follow infection ancestry. Mutations M1-4 generated by migration are marked I or S according to the migrating host type (AI or AS). Mutations M5-6 generated by non-transmission at birth are unlabelled. Host ancestry continues through non-transmission events. All ancestry terminates at migration events. In M2, M4 and M6, infection ancestry above the mutation is not indicated, as it is unrelated to ancestry below the mutation.

Fig. 6. Estimated probability densities for the hitting time of the boundary $W(t) = 1$ starting at $W(t_0) = 0.5$ with parameter values (a) $N_{CN} = \mu = 1.5$, $Ns_{N} = \sigma = 0$, $Nu_{N} = \theta = 1.3$ and (b) $N_{CN} = \mu = 4$, $Ns_{N} = \sigma = 1$, $Nu_{N} = \theta = 2$. 1500 realisations of each of three processes were simulated: the diffusion process (solid lines), the jump process (3) with $N = 200$ (dashed lines) and with $N = 400$ (dotted lines).