Available online at www.sciencedirect.com
SCIENCE DIRECT®
Theoretical Population Biology 68 (2005) 65–75
Theoretical Population Biology
www.elsevier.com/locate/ytpbi

Integrating genealogy and epidemiology: The ancestral infection and selection graph as a model for reconstructing host virus histories

David Welch[a], Geoff K. Nicholls[a,*], Allen Rodrigo[b], Wiremu Solomon[c]

[a]*Department of Mathematics, University of Auckland, New Zealand*
[b]*School of Biological Sciences, Allan Wilson Centre for Molecular Ecology and Evolution and Bioinformatics Institute, University of Auckland, New Zealand*
[c]*Department of Statistics, University of Auckland, New Zealand*

Received 17 November 2003
Available online 31 May 2005

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 and succeeding generations. The epidemic model allows a selective advantage for susceptible over infected hosts. The coupled host–virus ancestry of a sample of hosts 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 [1997. Theoret. Population Biol. 51, 210–237]. We prove this and discuss various special cases. We show that the inter-host viral genealogy is a scaled coalescent. Using simulations, we compare the viral genealogy under this model to earlier published models and investigate the estimatability of the selection and infectious contact rates. We use simulations to compare the persistence of the disease with the time to the ultimate ancestor.

Keywords: Epidemiology; Ancestral selection graph; Coalescent; Host–virus genealogy; Moran model; Population genetics

1. Introduction

Over the last decade, researchers have used viral evolutionary genetics to test hypotheses and/or estimate parameters relating to the epidemiology of disease (see, for example, Holmes et al., 1995; Twiddy et al., 2003; Pybus et al., 2000). In particular, the use of coalescent-based methods which rely on genealogical-based estimates of population parameters has been popular (Pybus et al., 2000). These methods use the genealogy of viruses, each obtained from a different host, to make inferences about host population dynamics. However, given that viruses are transmitted both horizontally (i.e., by contact) and vertically (from parent to child), it is not immediately apparent that naive coalescent-based methods are appropriate.

To explore this, we set up and analyse a model of the coupled host–virus genealogies of a panmictic host population. Our approach is to make appropriate modifications to the model of Moran (1958), an explicit model of mutation and selection. The genealogy process determined by the Moran model is well approximated by the ancestral selection graph (ASG) process of Krone and Neuhauser (1997). 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

*Corresponding author.
E-mail address: Nicholls@math.auckland.ac.nz (G.K. Nicholls).

0040-5809/$ - see front matter © 2005 Elsevier Inc. All rights reserved.
doi:10.1016/j.tpb.2005.03.003

λ_I. Susceptible individuals reproduce at a possibly higher rate λ_S. For some $s_N \geqslant 0$, 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.

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 corresponds 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 $\lambda_I c_N$, for some $c_N \geqslant 0$. The initiating host (the *initiator*) chooses another host (the *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.

Type before contact		Type after contact	
Initiator	Target	Initiator	Target
S	S	S	S
S	I	I	I
I	S	I	I
I	I	I	I

(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 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:

$$j \longrightarrow \begin{cases} j+1 \quad \text{at rate } \lambda_S j \dfrac{(N-j)}{N} + \lambda_I u_N (N-j) \\ \qquad \times \dfrac{(N-j)}{N} + \lambda_I \beta p \dfrac{(N-j)}{N}, \\ j-1 \quad \text{at rate } \lambda_I (1-u_N)(N-j)\dfrac{j}{N} + 2\lambda_I c_N j \\ \qquad \times \dfrac{(N-j)}{N} + \lambda_I \beta (1-p)\dfrac{j}{N}. \end{cases} \tag{3}$$

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 (1997), 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 μ, σ and θ so that

$$\left.\begin{aligned} |Nc_N - \mu| &\leqslant CN^{-\gamma} \\ |Ns_N - \sigma| &\leqslant CN^{-\gamma} \\ |Nu_N - \theta| &\leqslant CN^{-\gamma} \end{aligned}\right\} \quad \text{for all sufficiently large } N. \tag{4}$$

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

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

and diffusion $b(x) = x(1-x)$. 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)

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

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 terms like $\lambda_S u_N j^2/N$ in the rate for $j \to j-1$. 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 $\lambda_I c_N j(N-j)/N$. Direct mutation from susceptible to infected is represented by infected host migrants replacing susceptible hosts, a process with rate $\lambda_I \beta p j/N$. However, it follows from (4) that these three terms are $O(N)$. The models have the same diffusive limit for the proportion of susceptibles, up to a parameterisation. If, in the scaled Moran process, θ_I is

the mutation rate from A_I to A_S, θ_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_I 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 $\hat{\sigma} = 2\mu - \sigma \geqslant 0$. 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) (see Section 6 of Chapter 15), 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. However, the role of immigration in the model is not to impose a spurious persistence for the disease, but rather to allow host and viral ancestral lineages to terminate outside the model-population. 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 parameters μ, σ and θ 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. The presence of infected host in a sample indicates the process is not in equilibrium. This is discussed further in Section 2.

2. Graphical representations

A realisation of the infection process (3), acting in a population of hosts, can be represented via a percolation diagram. A realisation of the history of the infection for a sample of hosts is a subgraph of this diagram. In 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 (1997). These authors set up a percolation-diagram representation for the Moran model, and established the corresponding limiting ASG process.

2.1. The forward model

A realisation 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

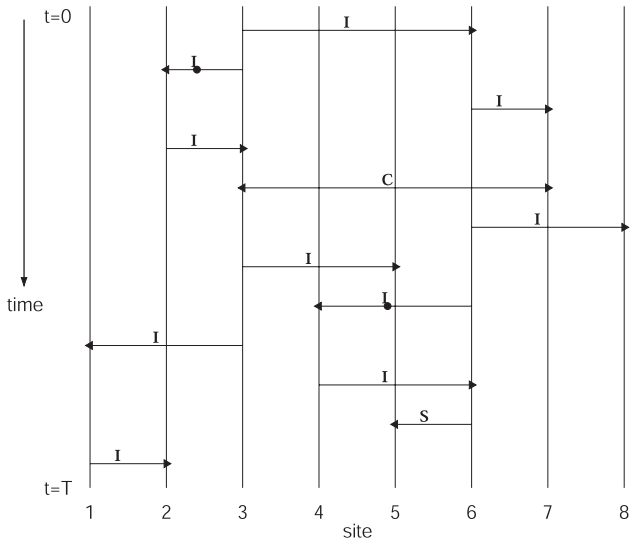

Fig. 1. A percolation diagram realisation of the infection model for $N = 8$. If at time $t = 0$, individuals at sites $\{1,3,5,7\}$ are susceptible and the rest infected then, propagating types according to the rules set out in the text, at $t = T$ individuals at sites $\{4,5,6\}$ are susceptible and the rest infected.

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 $\frac{\lambda_I(1+s_N+c_N)}{N}$. 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 $v < \frac{\lambda_I}{\lambda_I(1+s_N+c_N)}$ label the arrow I, if $v \in [\frac{\lambda_I}{\lambda_I(1+s_N+c_N)}, \frac{\lambda_S}{\lambda_I(1+s_N+c_N)})$ label the arrow S, otherwise $v \in [\frac{\lambda_S}{\lambda_I(1+s_N+c_N)}, 1]$ 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 λ_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 $n \leqslant N$ 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 (1997), we define a dual percolation process which gives us this information. A realisation of the dual percolation process is a subgraph of the percolation diagram realised by the forward process. The example pictured in Fig. 2 is derived from the realisation of the forward process depicted in Fig. 1.

We obtain a realisation 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 realisation 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.

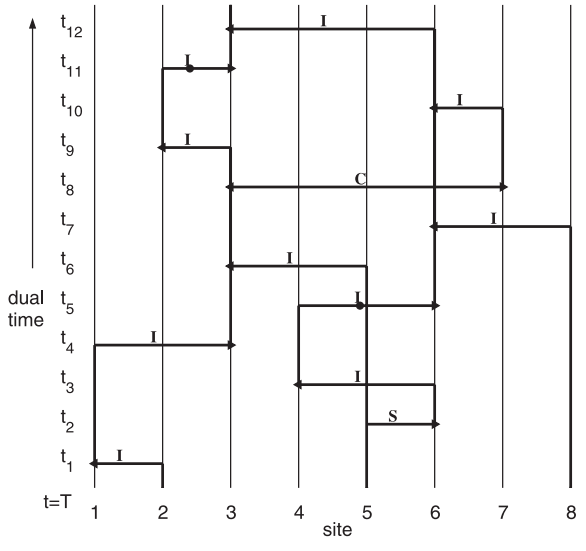

Fig. 2. The dual process for the infection model obtained by reversing the direction of time and the direction of the arrows in Fig. 1.

Consider site 2 in Fig. 2. At dual time t_1 the offspring of the individual at site 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 combined ancestral and infective history shown in bold in Fig. 2.

Because the forward process is reversible, and events on each site are realised 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 realised 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^N_{UA} denote the dual time of this joint ultimate ancestor.

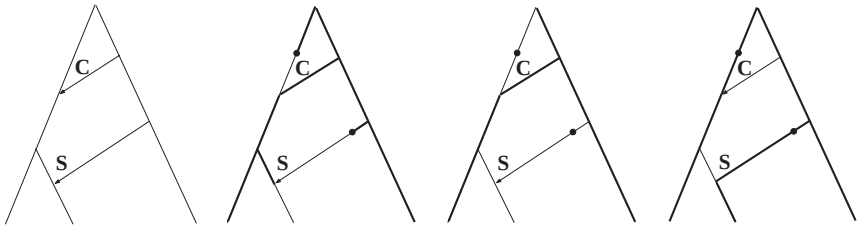

Fig. 3. The dual graph at left is derived from the dual percolation subgraph in bold in Fig. 2. Mutation events on the dual percolation subgraph in Fig. 2 determine corresponding events on the dual graph at left in Fig. 3. The individual at the top of the graph is infected. Tracing infection type down 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.

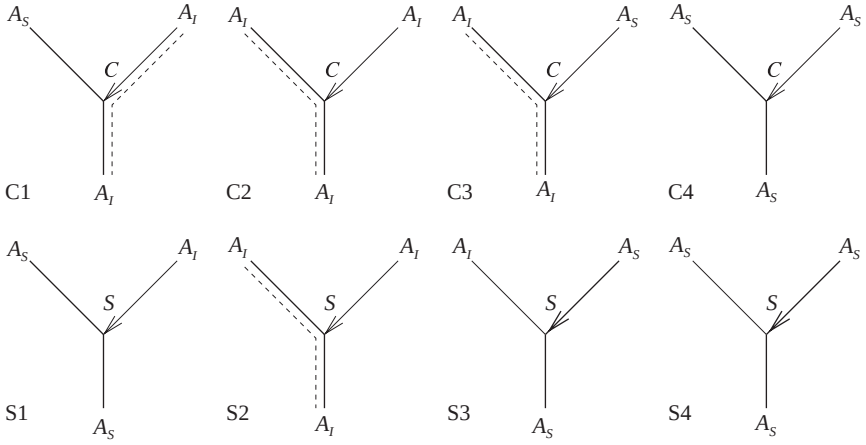

Fig. 4. Rules determining infection type and ancestry below a branching event at time t in the dual graph: infection types at dual times t^+ and t^- are indicated above and below the branching; an arrow indicates the incoming edge, the thick lines follow host ancestry, the dashed lines follow infection ancestry. Top row: rules for contact branchings. Bottom row: rules for selection branchings. In C4, S1, S3 and S4, no infection is present on the branching edge at t^- so no infection ancestry is indicated.

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, or 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. We call this cut-down realisation the dual graph. The dual graph in Fig. 3 summarises events shown in the bold subgraph of Fig. 2.

Denote by $\mathscr{G}_{N,n}$ the process realising dual graphs for samples of size n drawn from a population of size N. The graph on the left in Fig. 3 is a realisation of $\mathscr{G}_{8,3}$. Note that a realisation of $\mathscr{G}_{N,n}$ does not include mutation events, or details of infection type or ancestry. The mutation process on the dual percolation subgraph determines a mutation process $\mathscr{Y}_{N,n}$ on realisations of $\mathscr{G}_{N,n}$, i.e., on dual graphs.

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

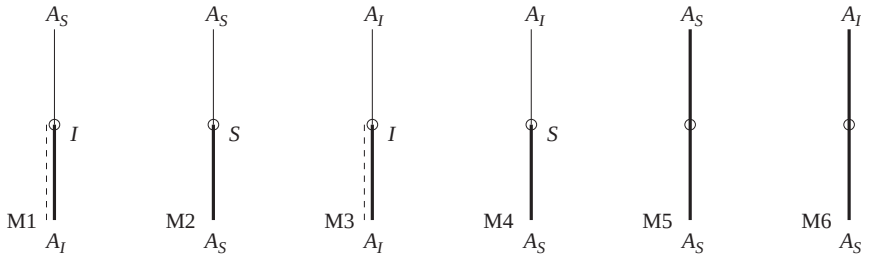

Fig. 5. Rules determining infection type and ancestry below a mutation event (indicated by a dot). The thick lines follow host ancestry and the dashed lines follow infection ancestry. Mutations M1–4 generated by migration are marked I or S according to the migrating host type (A_I or A_S). Mutations 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.

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, $\mathscr{G}_{N,n}$ and $\mathscr{Y}_{N,n}$, together converge in distribution, in the limit $N \to \infty$ defined in (4), to processes $\mathscr{G}_n$ and $\mathscr{Y}_n$. We refer to $\mathscr{G}_n$ as the ancestral infection and selection graph process (AISG-process) and to $\mathscr{Y}_n$ as the mutation process. Instantaneous rates for events in $\mathscr{G}_n$ 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 μ and S-branches at rate $\sigma/2$. A realisation of $\mathscr{G}_n$ starts at time t_0 with n lineages and terminates at the first time that k becomes one, t_{UA} say. The $\mathscr{Y}_n$-process realises 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 realisation of $\mathscr{G}_n$ and $\mathscr{Y}_n$ is determined according to the rules set out in Fig. 5. A detailed justification that $\mathscr{G}_n$ and $\mathscr{Y}_n$ are the limiting processes of the dual graph and mutation processes follows Krone and Neuhauser (1997) closely, and is therefore omitted. The following summary emphasises points of difference.

A realisation of $\mathscr{G}_{N,n}$ may contain events of a type which cannot be generated by $\mathscr{G}_n$. 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. Fig. 2 contains no collision events. An S-collision would have occurred if, for example, the S-arrow at dual time t_2 connected sites 1 and 5, instead of 5 and 6, since site 1 has a traced lineage at t_2 whereas site 6 does not. Similarly a C-collision would have occurred if the C-arrow at dual time t_8 connected sites 3 and 6 rather than 3 and 7. We stipulate that at each collision, a single fictional lineage is created and is allowed to evolve like all other lineages. If, at dual time t, graph process $\mathscr{G}_{N,n}$ has k lineages, each lineage encounters incoming 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 $\mathscr{G}_{N,n}$ coalesce at rate 1.

Consider a dual graph process, $\mathscr{G}^\star_{N,n}$, derived from a Moran process (as described in Section 1) with selective advantage $s^\star_N = 2c_N + s_N$. If $\mathscr{G}^\star_{N,n}$ has k lineages, each lineage branches at rate $(N-k)s^\star_N/2$ and encounters collisions at rate $ks^\star_N/2$. 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 $\mathscr{G}_{N,n}$ then $\mathscr{G}^\star_{N,n}$ and $\mathscr{G}_{N,n}$ are identical. Krone and Neuhauser (1997) show that in the limit $N \to \infty$, $\mathscr{G}^\star_{N,n}$ is the ASG $\mathscr{G}^\star_n$ with branching rate $\sigma^\star/2 = (2\mu+\sigma)/2$. The S- and C-branch (collision) labels can be independently imposed on $\mathscr{G}_{N,n}$ and $\mathscr{G}_n$, 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 $N \to \infty$, $\mathscr{G}_{N,n} = \mathscr{G}_n$.

The dual mutation process $\mathscr{Y}_{N,n}$ may include non-transmission events at the same dual time as coalescing events in $\mathscr{G}_{N,n}$. These simultaneous events almost surely do not occur in the $\mathscr{Y}_n$ process. Krone and Neuhauser (1997) show that the probability of these events

occurring in $\mathscr{Y}_{N,n}$ tends to 0 as $N \to \infty$. The immigration events in $\mathscr{Y}_{N,n}$ and $\mathscr{Y}_n$ occur as independent Poisson process along the edges of $\mathscr{G}_{N,n}$ and $\mathscr{G}_n$. The process $\mathscr{Y}_{N,n}$ is the combination of these two "mutation" processes (i.e., non-transmission and immigration). It follows from the above remarks that $\mathscr{Y}_{N,n}$ and $\mathscr{Y}_n$ are equivalent in the limit $N \to \infty$.

Notice that the features by which the ASG 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 graph process $\mathscr{G}_n$ resembles the two-locus ancestral graph process of Griffiths (1991), the complex disease process of Fearnhead (2003) and, in particular, 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 original ASG process of Krone and Neuhauser (1997) 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 AIG process are of two types, accounting 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 AIG 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. Properties of the embedded genealogies

The $\mathscr{G}_n$ and $\mathscr{Y}_n$ processes determine a marginal distribution for the genealogy of the virus. This marginal distribution is of interest in its own right, as a model of inter-host viral genealogy. In this section we show that the marginal viral genealogy is a coalescent process with population size given by a diffusion. The intra-host viral genealogy may be simulated directly, without the need to simulate host genealogies. In future work we will be concerned with fitting the joint host–virus model to sequence data, and in particular, estimating the contact and selection parameters, μ and σ. Neuhauser and Krone (1997) have shown that genealogies determined from the ASG by the simplest two-type symmetric substitution process are rather insensitive to the selection parameter, making estimation difficult. We present simulation results which show that the μ-dependence of genealogies in the AISG is somewhat stronger than the σ-dependence.

Inter-host viral genealogies have been modelled by previous authors (see for example Holmes et al., 1995; Pybus et al., 2000) using the Kingman coalescent. In these models, the effective population size is $N_e = N(1 - x(t))$, where N is the total number of hosts and $x(t)$ is the proportion susceptible at time t. The behaviour of N_e back in time is variously modelled as constant, growing exponentially (Holmes et al., 1995) or piecewise constant to approximate arbitrary continuous change (Pybus et al., 2000). In the AISG, the size of the infected population is $N(1 - W(t))$ with $W(t)$, the proportion of susceptible hosts in the population at time t, defined in Section 1.

Theorem 3.1. *If at time t the population fraction of susceptibles is $W(t) = x$ then, in dual time, and in units of N generations,*

(1) *each pair of lineages in the genealogy of the virus coalesces at instantaneous rate* $(1 - x)^{-1}$,
(2) *each pair of infected lineages in the host genealogy coalesces at instantaneous rate* $(1 - x)^{-1}$ *and each pair of susceptible lineages coalesces at instantaneous rate* x^{-1}.

Proof. The result is essentially identical to that given in Kaplan et al. (1988) for the marginal coalescent of a selected allele. Consider the dual percolation process. Denote by j the number of susceptibles in the host population at dual time t. Take two infected individuals at random from the set of pairs of infected individuals at $t = 0$ and trace their viral lineages back to dual time t. Suppose they do not coalesce in $[0, t)$. The two viral lineages sampled at dual time t by this procedure are a random draw from the set of pairs of viral lineages present at time t. In forward time an infected lineage generates infected individuals through birth and contact at rate $\lambda_I(1 - u_N) + \lambda_I c_N$ and, since it may choose its own site, the probability that any one of the $N - j$ infected individuals in the next generation is the offspring of that branching event is $1/(N - j)$. It follows that in dual time a pair of viral lineages in the dual percolation process coalesce at instantaneous rate $2\lambda_I(1-u_N)/(N-j)+2\lambda_I c_N/(N-j)$. Taking $N \to \infty$ subject to (4), the second term vanishes and the first term, $2\lambda_I(1 - u_N)/(N - j) = 2\frac{N}{2}(1 - \frac{\theta}{N})\frac{1}{(1-x)N} = \frac{N-\theta}{(1-x)N} \to \frac{1}{1-x}$. The proof of (2) is similar. □

Theorem 3.1 tells us how to simulate viral genealogies backwards in time. Denote by $\tilde{W}$ the reversed diffusion $\tilde{W}(t) = W(-t)$. If both $W = 0$ and 1 are not absorbing, $\tilde{W}(t)$ is identical in distribution to $W(t)$. If one or both of $W = 0$ and 1 are absorbing, the reverse process starting from $\tilde{W} \in (0, 1)$ is identical in distribution to W

in $(0,1)$. In these cases the process $\tilde{W}$ is killed at $W=1$ or absorbed at $W=0$ as appropriate. We are assuming the "return processes" (see Tavaré, 1979) from absorbing boundary $W=0$ adds one susceptible to an all-infected population and from absorbing boundary $W=1$ adds one infected individual. When both $W=0$ and 1 boundaries are absorbing it may be realistic to condition on entry from $W=1$. Coop and Griffiths (2004) describe and make statistical inference for a problem of this kind.

In order to simulate directly the marginal viral genealogy given an initial population fraction $W(0)=x_0$ of susceptibles, simulate $\tilde{W}(t)$ by simulating $W(t)=x(t)$ from $W(0)=x_0$ in forward time from $t=0$ and then setting $\tilde{W}(-t)=x(t)$. Simulate coalescence times in the viral genealogy conditional on this realization of $\tilde{W}(t)$ at increasing positive dual times. The simulation stops when the number of viral lineages reaches 1. This must occur before the $\tilde{W}$ process is killed in the case $W=1$ is absorbing, as the infected population shrinks to zero. Notice that the marginal viral genealogy need not be a connected graph when immigration with $p\beta>0$ is present. For simulations, we approximate $W(t)$ by the Moran process Y described by rates (3) with $\beta=0$ and N large.

The host genealogy is embedded within a realization of the ASG process. In order to simulate a host genealogy, we could first simulate a realization g_0 of the ASG with branching rate σ, and then simulate contact branches conditioned on g_0. The result is a realization, $\mathscr{G}_n=g_1$ say, of the AISG which contains g_0 as a subgraph. Host and viral ancestry is determined from infection types entering vertices of g_1, which are in turn realized by simulating $\mathscr{Y}_n$ over g_1. Now, referring to Fig. 4 C1–4, host lineages do not follow contact branches, so the host genealogy is itself a subgraph of g_0. The effect of contact, like the effect of non-transmission, is to determine the path taken by the host genealogy through the ASG. These observations do not lead to a direct simulation scheme for the marginal host genealogy, since host infection status is not available at all dual times.

We simulated marginal viral trees in order to compare their distribution with that of a standard coalescent process, K_n say. We estimated the distribution of inter-coalescence times for an initial sample of size $n=10$, and initial frequency of susceptibles, $x_0=0.5$, simulating genealogies using the prescription above. We compare these times with inter-coalescence times for viral genealogies in populations with a constant proportion $x(t)=0.5$ of infecteds. Results for parameter values $\mu=2$, $\sigma=1$ and $\theta=1.5$ are summarised in Fig. 6. For these parameter values we estimate $E_{G_{10}}[T_{MRCVA}]=0.51(1)$ under the diffusing model, and calculate $E_{K_{10}}[T_{MRCVA}]=0.90$ under the model with constant (infected) population size $N(1-x)=N/2$.

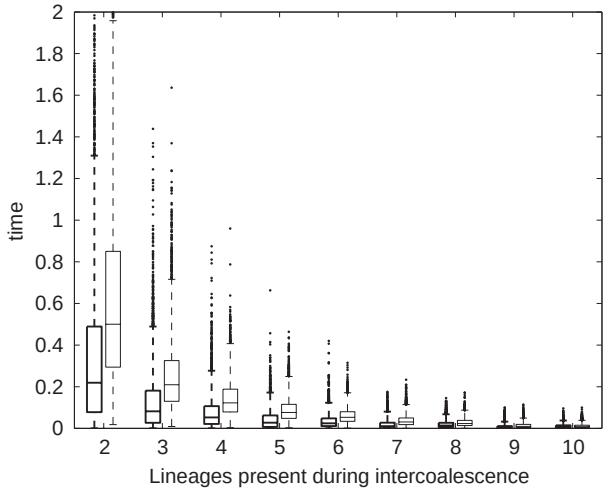

Fig. 6. Distributions of inter-coalescence times are shown for initial sample size $n=10$ and initial susceptible fraction $x_0=0.5$ for (thin lines, right-hand box of each pair) a constant population fraction of susceptibles and (thick lines, left-hand box) a diffusing population fraction of susceptibles, with $\mu=2$, $\sigma=1$ and $\theta=1.5$. 4000 viral genealogies were simulated in each case using the Moran model approximation with $N=2000$ and the marginal simulation scheme of Section 3. Dots are samples lying outside 2.5 times the inter-quartile range of the median sample.

Diffusion-model variance is larger, relative to its mean: $\mathrm{var}_{G_{10}}[T_{MRCVA}]=0.23$ for the diffusing case and $\mathrm{var}_{K_{10}}[T_{MRCVA}]=0.29$ in the constant case. Patterson (2004) shows that if the proportion of susceptibles is neutrally diffusing (i.e., has zero drift everywhere and diffusion coefficient $x(1-x)$), the expected coalescence times are the same as that under the constant proportion model. Patterson (2004) gives a coalescent simulation algorithm which avoids the need to simulate $\tilde{W}(t)$ at times which are not coalescent times. Our simulations show that we cannot expect equivalent results to hold in our setting.

Finally in this section, we use simulations to investigate the sensitivity of the embedded genealogies to changes in parameters σ and μ for fixed θ. Such variation plays an important role in parameter estimation. We look at the case where there are k leaves all of which are infected (results are given for $k=10$). We fix $n\geqslant k$ ($n=20$ below) and simulate multiple realizations of the AISG and type processes $(\mathscr{G}_n,\mathscr{Y}_n)$. The host and viral genealogies of a randomly chosen subset of k infected leaves are then extracted. If the realization of $(\mathscr{G}_n,\mathscr{Y}_n)$ lacks k infected leaves the sample is discarded. This thinning procedure simulates the host and viral genealogies of k individuals selected at random from the infected host population. It would be more efficient, but less convenient, to add leaves and make conditional simulation until k infected leaves were generated.

Mean surfaces for various statistics of interest obtained from simulations with fixed $\theta=1$ are

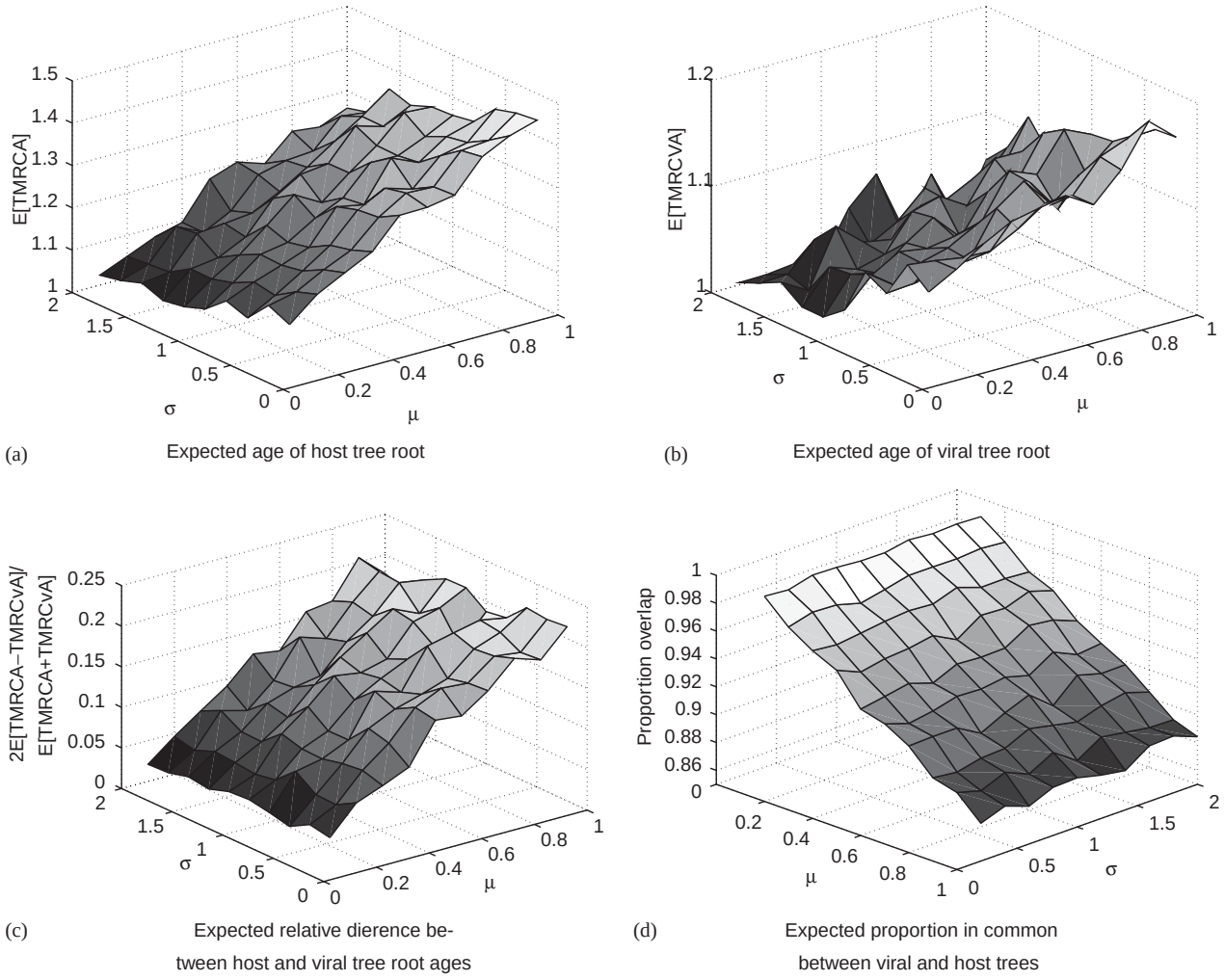

Fig. 7. 6000 AISG graphs with $n = 20$ leaves were simulated for each of 100 (θ, σ, μ)-values in $\{(\theta, \sigma, \mu) : \theta = 1, 0.2 \leqslant \sigma \leqslant 2, 0.1 \leqslant \mu \leqslant 1\}$. Depending on the parameter values, this produced, at each (θ, σ, μ)-value, between 763 and 2593 host and viral genealogies with $k = 10$ infected leaves. (a) Expected age of host tree root; (b) expected age of viral tree root; (c) expected relative difference between host and viral tree root ages; (d) expected proportion in common between viral and host trees.

presented in Fig. 7. Embedded genealogies are more sensitive to changes in the contact parameter μ than to changes in the selection parameter σ. The statistic in Fig. 7(d) is calculated as follows. Denote by $L(E_h^g)$ and $L(E_v^g)$ the summed length of edges in the host and viral genealogies, and by $L(E_{h,v}^g)$ the summed length of edges, or parts of edges present in both genealogies. The statistic $E[2L(E_{h,v}^g)/(L(E_v^g) + L(E_h^g))]$ plotted in Fig. 7(d) is a natural measure of the distance between the host and viral genealogies.

Neuhauser and Krone (1997) took a sample of 30 ASG genealogies subject to selection of $\sigma = 2$ and considered the hypothesis that the sample was drawn from a neutral model. They constructed a test statistic based on root times and found that the neutral model could not be rejected. In our simulations large sampling variances overwhelm small variations in mean. For example, in Fig. 7(d), the standard deviation of $2L(E_{h,v}^g)/(L(E_v^g) + L(E_h^g))$ is about 0.13, which is similar in magnitude to the variation across the surface. Accurate estimation of μ or σ will require large sample sizes of infected hosts. New data types may help. Sequence data gathered at different time points and data measuring $x(t)$ directly entail new estimation schemes, but will be informative.

4. Special cases

The infection model and AISG-process described in Section 1 belong to a class of models for the joint ancestry of host and parasite. In this section, we consider a model with no immigration and a model with an asymmetric contact process. The immigration process of Section 1 terminates host ancestry at both infected and susceptible host immigration events. A viral immigration (for example, a zoonosis) at rate $\lambda_I \beta^\star$

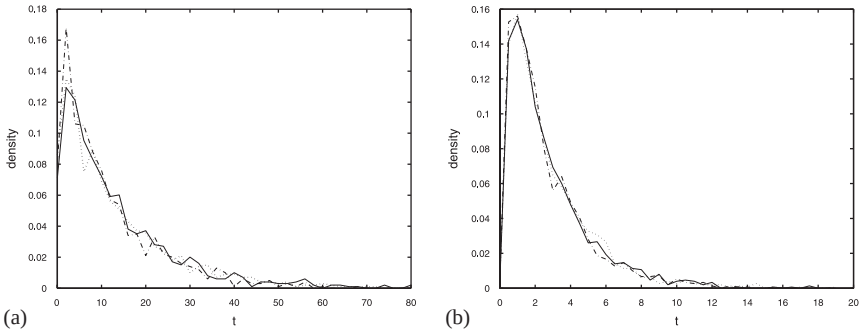

Fig. 8. Estimated probability densities for the hitting time of the boundary $W(t) = 1$ starting at $W(t_0) = 0.5$ with parameter values (a) $Nc_N = \mu = 4$, $Ns_N = \sigma = 1$, $Nu_N = \theta = 2$ and (b) $Nc_N = \mu = 5$, $Ns_N = \sigma = 5$, $Nu_N = \theta = 3$. 1500 realisations of each of three processes were simulated: (solid lines) $W(t)$ the diffusion process of (5) simulated with Gaussian increments, and the jump process (3) with $N = 200$ (dashed lines) and with $N = 400$ (dotted lines).

which does not terminate host ancestry is quite tractable.

When there is infection in the population, but no immigration, the infection descends from the joint ultimate ancestor, which has type A_I. For this scenario to be physically relevant, the infection must remain within the population for a time comparable to the time to the joint ultimate ancestor, t_{UA}, and increases as $2\mu + \sigma$ increases. The expected time that the infection persists is the expected time to hit the absorbing boundary at $W(t) = 1$ and increases as $2\mu - \sigma$ increases.

If $\beta = 0$, the diffusion process $W(t)$ describing the proportion of susceptibles in the population has drift $a(x) = (\theta(1-x) - (2\mu - \sigma)x(1-x))/2$. This has zeros at $x = 1$ and $x = \theta/(2\mu - \sigma)$. Since $x \in [0,1]$, $a(x) > 0$ when $\theta/(2\mu - \sigma) < 0$ or > 1 and, in these cases, the process is attracted to the absorbing boundary at $W = 1$. The process remains on the interior of the space longest when $\theta/(2\mu - \sigma) \in (0,1)$ and $2\mu - \sigma$ is large. In this case, $\theta/(2\mu - \sigma)$ 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 persistence time of the epidemic can be compared with the expected time to the joint ultimate ancestor, t_{UA}, which can be calculated from Krone and Neuhauser (1997) via

$$E_n[t_{UA}] = 2\left(1 - \frac{1}{n}\right) + 2\sum_{r=1}^{n-1} \frac{1}{r(r+1)} \frac{e^{2\mu+\sigma}}{(2\mu+\sigma)^{r+1}} \times \int_0^{\sigma} t^{r+1} e^{-t} dt, \tag{7}$$

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.

Using simulation we estimate the length of time that an infection persists. Fig. 8(a) summarizes a simulation study for parameter values $\mu = 4$, $\sigma = 1$, $\theta = 2$ and $W(0) = 0.5$. These values gave an expected hitting time equal $13N$ generations. Fig. 8(b) shows results for $\mu = 5$, $\sigma = 5$, $\theta = 3$ and $W(0) = 0.5$. The expected hitting time was approximately $2.8N$ generations. For Fig. 8(a), $2\mu + \sigma = 9$, and (7) gives $E_2[t_{UA}] \approx 200N$ generations, while for Fig. 8(b), $2\mu + \sigma = 15$, and we find $E_2[t_{UA}] \approx 30,000N$ generations. Either immigration plays an important role, or the diffusive limit based on scaling relations (4) is not physically relevant. When the latter applies it may be necessary to fit data to the dual processes $\mathscr{G}_{N,n}$ and $\mathscr{Y}_{N,n}$.

When we model a virus which causes the host to act in a way that spreads the virus, for example, causing the host to sneeze, contact between hosts is not symmetric. The contact scheme defined in Table 2 assumes that all hosts initiate contact and that the parasite may be passed equally from the initiator to the target or *vice versa*. When infected hosts alone initiate contact then in (3), the second term in the rate for decrease is halved to $\lambda_I c_N j(N-j)/N$. The AISG is unchanged except that the C-branching rate is halved to $\mu k/2$.

5. Discussion

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 rate of immigration of foreign hosts or 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 interpretation and details of the AISG differ from the

ancestral selection graph, the graphical structures are identical. 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 investigated the marginal distributions of the embedded host and viral genealogies. We showed that the viral genealogy can be simulated independently of the full AISG by appropriate scaling of a standard coalescent and that the host genealogy is embedded in a subgraph of the AISG obtained by ignoring the contact branches. These results, and our simulation studies, show that both genealogies are sensitive to changes in any of the contact, selection and non-transmission parameters.

Returning to the issue of whether it is appropriate to use simple coalescent-based models of viral genealogies to make inferences about host population dynamics, our simulations show that horizontal and vertical transmission rates can influence population estimates. The simplest estimate we can derive—an estimate of the infected population size—using the simplest estimator (the time to the most recent common ancestor) can be biased—this is the message of Fig. 6. Direct use of the coalescent to model inter-host viral genealogies may be inappropriate. We may be better served by an integrated model of viral and host genealogies. In this regard, it seems likely that for real data, the efficiency of our estimates will increase substantially if we had information on both the viral and host genealogies for large sample sizes, for serial data including sequences and direct estimates of the infected population fraction gathered at intervals from the population. This represents a future direction for our work.

References

Coop, G., Griffiths, R.C., 2004. Ancestral inference on gene trees under selection. Theoret. Population Biol. 66, 219–232.

Donnelly, P., Kurtz, T., 1999. Genealogical processes for Fleming–Voit models with selection and recombination. Ann. Appl. Probab. 9, 1091–1148.

Ewens, W.J., 1979. Mathematical Population Genetics. Springer, Berlin.

Fearnhead, P., 2003. Ancestral processes for non-neutral models of complex diseases. Theoret. Population Biol. 63, 115–130.

Griffiths, R.C., 1991. The two-locus ancestral graph, 1989. In: Basewa, I.V., Taylor, R.L. (Eds.), Selected Proceedings of the Symposium on Applied Probability, IMS Lecture Notes, vol. 18, Institute of Mathematical Statistics, Sheffield, pp. 100–117.

Holmes, E.C., Nee, S., Rambaut, A., Garnett, G.P., Harvey, P.H., 1995. Revealing the history of infectious disease epidemics through phylogenetic trees. Philos. Trans. Biol. Sci. 347, 33–40.

Kaplan, N.L., Darden, T., Hudson, R.R., 1988. The coalescent process in models with selection. Genetics 120, 818–829.

Karlin, S., Taylor, H.M., 1981. A Second Course in Stochastic Processes. Academic Press, New York.

Krone, S.M., Neuhauser, C., 1997. Ancestral processes with selection. Theoret. Population Biol. 51, 210–237.

Moran, P.A.P., 1958. Proc. Cambridge Philos. Soc. 54, 60–72.

Neuhauser, C., Krone, S.M., 1997. The genealogy of samples in models with selection. Genetics 145, 519–534.

Patterson, N.J., 2004. How old is the most recent ancestor of two copies of an allele? In preparation.

Pybus, O.G., Rambaut, A., Harvey, P.H., 2000. An integrated framework for the inference of viral population histories from reconstructed genealogies. Genetics 155, 1429–1437.

Tavaré, S., 1979. Dual diffusions, killed diffusions, and the age distribution problem in population genetics. Theoret. Population Biol. 16, 253–265.

Twiddy, S.S., Pybus, O.G., Holmes, E.C., 2003. Comparative population dynamics of the mosquito-borne flaviviruses. Infec. Gen. Evol. 3, 87–95.