Center for Studies in Demography and Ecology

Incorporating Human Mixing Patterns into HIV 1 Population Genetics

by

Steven M. Goodreau
University of Washington

Incorporating human mixing patterns into HIV-1 population genetics

Steven M. Goodreau

Center for AIDS and STD, Center for Statistics and the Social Sciences

University of Washington, Seattle WA 98195

Abstract:

Geneticists seeking to understand the evolution of HIV-1 among populations of human hosts generally apply models that assume the hosts belong to a single panmictic population. Social science research indicates that the network patterns over which HIV-1 is spread are far from random, but the effect of these patterns on the genetic diversity of HIV-1 and other sexually transmitted pathogens has yet to be examined. This study uses a graph-theoretic framework known as exponential random graph modeling (ERGM) to simulate seven types of dynamic partnership networks resembling those observed in real human populations and to simulate HIV transmission and mutation in these networks. The means of the resulting mismatch distributions reveal that effective viral population size for the structured mixing populations is 10-24% higher than that observed for viral populations of the same size generated in a randomly mixing human population. The author discusses implications for HIV-1 phylogenetics along with the potential for ERGM to provide a general framework for addressing these issues.

Understanding the nature of genetic variation in HIV-1 is a crucial public health issue, since the high mutation rate of this retrovirus leads to myriad genetic forms that may exhibit different properties of infectivity, virulence, or susceptibility to vaccines and medicines. Geneticists have spent considerable effort both in describing the distribution of this variation globally or in a given community, and in uncovering the functional differences these forms create. These tasks are complicated by the fact that the object of study forms a meta-population with different evolutionary forces operating at two levels. That is, each HIV-1-infected host contains a local pathogen population (often called a "quasi-species" in the viral literature), and these populations are then linked together epidemiologically and phylogenetically into a single connected meta-population. Genetic diversity among viral particles within each host is a function of the forces population geneticists are used to thinking of: mutation, drift and selection, both host-mediated and otherwise. The forces shaping genetic diversity among the components of this meta-population include not only these more traditional population genetic concepts, but also the patterns of host behavior that allowed for pathogen transmission among the hosts. Considerable work on the genetics of metapopulations already exists (e.g. Hanski and Gilpin 1997); however, the metapopulation dynamics of HIV-1 are unique in a number of ways (new demes founded at rates determined by patterns of human sexual behavior, demes completely isolated once founded, guaranteed extinction of demes ~5-20 years after their founding) and require an exploration that incorporates these unique features.

Much work has now been done to relate patterns of HIV-1 genetic diversity to population history for host populations that are either constant or growing in some parametric trajectory (e.g. Grassly et al. 1999, Holmes et al. 1999, Pybus et al. 2000).

However, such work almost always assumes that the mixing among these hosts is completely random, even though common sense and a large body of social science research make it clear that for sexually transmitted infections this is never the case (e.g. Laumann et al. 1994). Such substructure can easily be imagined to introduce irregularities into the growth patterns of the epidemic and consequently into the genetic diversity within that epidemic. For instance, a population that is strongly divided into sub-populations should exhibit a punctuated equilibrium in infection rates as the virus spreads rapidly within sub-populations but then waits a relatively long time to pass into another sub-population (Hanski and Gilpin 1997).

In this paper I present a graph-theoretic framework for quantifying patterns of non-random ("structured") host mixing and incorporating them into population genetics. This framework involves a class of probability models, known as exponential random graph models, for the stochastic microsimulation of networks of social relationships. I apply these models to the simulation of populations undergoing six types of structured mixing (100 populations per mixing type), and simulate HIV-1 transmission and evolution within these populations. I then compare the genetic variation resulting from the various forms of mixing networks to that resulting from a panmictic population by sampling one 500-nucleotide viral sequence from each infected host and examining the pairwise genetic distances between sampled sequences in each population. The selection of specific model parameterizations to examine within this framework is informed by research on community-level sexual network structure from the social sciences.

BACKGROUND

HIV-1 phylogenetics: Applications of phylogenetic methods to HIV-1 were at first largely focused on determining patterns of relatedness within some subsection of the epidemic by elucidating phylogenetic relationships among a sample of sequences. The patterns of relations reconstructed ranged from single infection chains (e.g. Ou et al. 1992 for a dental practice in Florida), through intermediate population and geographic levels (e.g. Leigh Brown et al. 1997 for relations among sequences in six cities in the British Isles) to global and multi-species relations (e.g. McCutchan et al. 1992 for relationships among the major geographical subtypes of HIV-1).

Recent work has focused on reconstructing population dynamics of the virus, most commonly the size and growth rate of some section of the epidemic through time. This literature has used increasingly sophisticated models and techniques to address the effective size of the viral pool among hosts for different strains of the virus in different regions and times. Grassly et al. (1999) used mismatch distributions to compare the population history of HIV-1 subtypes *A* and *B*. Holmes et al. (1999) used a technique developed earlier (Nee et al. 1994 and 1995) to compare the reconstructed number of phylogenetic lineages through time to that expected under a variety of population history models, including constant population size, linear growth, and exponential growth. In Pybus et al. (2000) they extended this method to include the generation of maximum likelihood estimates of the effective population size at all points in the past. In general, however, although these models have expanded their methods of analysis and the range of population growth patterns that they cover, they all continue to assume that mating is

random. One exception is Grassly et al. (1999), who compare their global subtype mismatch data to models of population panmixis and binary subdivision. They found that at this global level, the improved fit of the data to the subdivision model was not sufficiently large to warrant the increase in model complexity.

These researchers were not trying to determine directly the actual census size of the host population for a given strain. Central to their approach is the concept of effective population size (N_e), the mathematical construct referring to the hypothetical population (known as a Wright-Fisher population) that would yield an observed mean level of genetic diversity if all members had an equal chance of contributing offspring to the next generation. "Equal chance" requires a lack of strong selection, and implies certain outcomes—among them, that the number of offspring left in a subsequent generation by each member of the prior generation follows a Poisson distribution, and that there is no correlation between the number of offspring left by successive generations in the same lineage. Genetic diversity in this type of population is generated in well-understood ways.

Of course few real populations fit all of the assumptions of the Wright-Fisher model. The value of this framework comes from the numerous formulas that convert effective population size into census population size for different violations of the model assumptions. This provides a common metric for comparing populations and for testing a variety of hypotheses about their demographic history or their exposure to selection. Unfortunately, the complex partnership patterns we see in human sexual/drug-sharing networks do not fit neatly into any of the existing extensions. For instance, any level of assortative mixing by partner number (that is, highly active people tending to have highly active partners) means that there will be a correlation among offspring number in

successive generations of the virus by lineage, purely for network reasons. Grassly et al. (1999) have suggested that it is reasonable to assume a lack of correlation among generational offspring number; the justification for this statement was that there is no known biological difference among strains in terms of infectivity. There are, however, very clear behavioral differences operating among the hosts carrying different lineages, and evolutionary outcomes for HIV-1 are dependent jointly on the biology of the virus and the behavior of the host. Other network patterns may also lead to relationships among lineages that have not been previously addressed. This implies that we have little means for relating effective population size and census population size in these populations, and thus for conducting investigations of HIV-1 evolutionary processes that require such estimates. My goal in this paper will be to begin remedying this situation by introducing a graph theoretic approach to modeling sexual partnership patterns. This approach is described in the next section.

The genetic outcomes to be examined from these populations include a single 500-nucleotide viral genetic sequence for each infected actor in each population. I will use the mean of the pairwise genetic distances between these samples from each population to compare effective viral population sizes resulting the various types of structured mixing to random mixing. Although this summary measure of genetic diversity is less powerful than those based on complete coalescent methods (Felsenstein 1992), the latter require populations to be large relative to the samples drawn from them for study, a requirement that is not fulfilled in this case.

Coalescence theory states that for a haploid organism, effective population size can be estimated from the mutation rate and the mean genetic distance between any two samples in a population:

$$N_e = \frac{\delta}{2\mu} \qquad \text{[Eq. 1]}$$

where μ is the generational mutation rate and δ is the mean genetic distance (i.e. for a population of size *n*, the fraction of sites at which two sampled sequences differ averaged over the $\binom{n}{2}$ pairs). Equation 1 is based on the observation of Watterman (1975) that the expected amount of time in the past that two sequences in a Wright-Fisher population of size *N* have their most recent common ancestor (MRCA) is *N* generations. Thus between any two contemporary sequences there is an expected lineage time on which to accumulate mutations of 2*N* (the distance from MRCA to the first sequence plus the distance from MRCA to the second sequence). The expected genetic distance between two samples δ is thus simply a product of amount of mutation time and the mutation rate, or $2N\mu$. N_e can then be substituted for *N* since the effective size of a population is simply the size of a Wright-Fisher population with the same δ.

The distribution of the mismatch δ can in turn be related to population history for certain patterns of changing population size. In a stable panmictic population, for instance, the mismatch distribution is generally ragged. For a population that has been growing exponentially, Slatkin and Hudson (1991) show that when $N_0 r \gg 1$ the mismatch distribution is Poisson with a mean of:

$$\delta = \frac{2\mu\left[\ln\left(N_0 r\right) - \gamma\right]}{r} \qquad \text{[Eq. 2]}$$

where N_0 is the ending population size, r is the exponential growth rate per generation, and γ is Euler's constant, 0.577. Unfortunately, no such formulas have been worked out to link δ to viral population dynamics resulting from hosts with complex mixing patterns.

Any attempt to remedy this situation must consider the elaborate intrahost dynamics of the virus as well as interhost. Because of the virus's high mutation rate and a type of recombination that it undergoes during replication, each infected person harbors an entire quasispecies of HIV-1; Vartanian et al. (1992) estimate that among ~10^{10} infected host cells during clinical latency there exist on the order of 10^8 different viral genetic sequences. This set of variants will be constantly changing through drift, mutation and selection as viral particles die and new ones appear. Towards the end of infection, pairwise genetic differences for HIV-1 sequences within a single host may be as high as 10% (Ahmad et al. 1995). All of this would seem to preclude any hope of understanding relationships among viral sequences in different hosts.

Despite these elaborate intrahost dynamics, researchers have been successful in applying the molecular clock to HIV-1. For instance, Leitner et al. (1996) used a unique epidemiological case—a chain of nine individuals in Sweden linked together in a transmission cluster with known transmission times—to demonstrate molecular clock-like behavior over a period of 14 years. There appear to be many reasons why the epidemiological molecular clock is more robust to intrahost dynamics than one would initially expect: only a small fraction of infected cells are replicating, each new individual

seems to be infected with only a single strain, recombination prevents strong selective sweeps, and many of the selection pressures are diversifying rather than purifying. This does not mean that intrahost dynamics can simply be ignored, however, and the basic elements will be incorporated into the model of interhost phylogenetics used in this paper, as shown below.

Network epidemiology: The methods used here to model transmission are drawn from social network analysis, an outgrowth of both graph theory and social theory in which analysis focuses not only on a set of social agents but on the relationship ties between pairs of those agents as well. Within this framework I focus on a probability model class known as exponential random graph modeling, or p* (*p-star*) modeling, which was first proposed in the spatial statistics literature (Besag 1974), expanded upon by Frank and Strauss (1986) and Strauss and Ikeda (1990) and introduced to social network analysis by Wasserman and Pattison (1996). This approach derives a model for partnership formation by defining probabilities for each possible graph (or network; the terms are used interchangeably) of size n (i.e. containing n actors). Let x_{ij} represent the value of the tie between nodes i and j; if, as here, relationships are binary ($x_{ij} = 1$ if a tie is present or 0 if a tie is absent) and non-directed ($x_{ij} = x_{ji} \; \forall \; i,j$), then a graph x is defined by its $\binom{n}{2}$ tie values $x = \{x_{1,2},\ x_{1,3}, x_{1,4} \ldots . x_{2,3} \ldots . x_{n-1,n}\}$. In its general form, the model represents the marginal probability that a random graph X of size n will take on a value x as:

$$P(X = x) = \frac{\exp\left(\theta^T z(x)\right)}{c(\theta)} \qquad \text{[Eq. 3]}$$

where $z(x)$ is a vector of network statistics, θ is a vector of parameters, and $c(\theta)$ is a normalizing constant to ensure that the probabilities sum to one over all graphs of node size n. Examples of commonly used z statistics include the number of ties in the graph, the number of nodes with a certain number of ties over the observed time period, or the number of triads (sets of three nodes who possess all three pairwise ties).

This representation is so general as to include any possible probability model based on network statistics (Besag 1974); some guidance is thus required in choosing specific parameterizations. This can be provided by the growing number of studies that have collected some form of network data on sexual relationships, as well as modeling work showing the effect of network structure on HIV-1 transmission patterns. These include (among others) studies in Colorado Springs, Colorado (Klovdahl et al. 1994); New York, New York (Martin 1987, Morris and Dean 1994); Seattle, Washington (Garnett et al. 1996); Brooklyn, New York (Friedman et al. 1997); suburban Atlanta, Georgia (Rothenberg et al. 1998); Iceland (Haraldsdottir et al. 1992); Nigeria (Orubuloye et al. 1992); and Thailand (Morris et al. 1996). The common observation of both empirical studies and modeling efforts is that network structure matters for epidemic outcomes above and beyond the levels of activity. Assortative mixing by social attributes (race, age, location, occupation) is commonly observed and has been shown through modeling to affect the dynamics of disease spread. For instance, Morris and Dean (1994) simulate observed assortative mixing patterns by age among gay men in New York City and demonstrate that this bias, if maintained in the long-term, should reduce prevalence among young gay men 40% below what it would be if mixing were random. The

existence of a "core group" that is not only highly active but preferentially mixes with other highly active people can also be important (Garnett et al. 1996), although perhaps more so for short-term curable STDs than for HIV. Watts and May (1992) and Morris and Kretzschmar (1997) demonstrated that the timing of partnerships (concurrent vs. serial) is a strong determinant of prevalence. Another important pattern includes the existence of "bridges," or individuals who serve as epidemiological links between groups that otherwise would have no interaction, as men do when they have both commercial and non-commercial female sex partners (Morris et al. 1996). Unfortunately, none of these studies have sequenced HIV-1 sequences from seropositive subjects, allowing for a simultaneous examination of network structure and phylogenetics in a real population. Instead, I draw upon the qualitative observations of these studies to determine the types of mixing patterns to examine.

METHODS

From the above literature, I chose seven sets of mixing rules to examine: one panmictic population (*random model*), four populations divided into equally active subgroups with internal preferential mixing (*assortative models*), one population with a small highly active subgroup (*core model*), and one *bridge model*. This last population consists of husbands, wives, and non-married females; each husband/wife pair maintains their relationship throughout the course of the simulation, while husbands may have simultaneous ties with non-married females. Details of the individual models are listed in

Table 1, while the z statistics and associated θ parameters necessary to create these models are listed in Table 2. These values were calculated using a likelihood approach described by Strauss and Ikeda (1990)[1].

Each model contains 200 actors sharing an expected 200 partnerships at any moment in time. Populations are small since computing needs for network simulation generally scale with the square of the population size, a continuing limitation on network-based methods. Partnerships contain two actors, so 200 partnerships translates into an average of two partnerships per person at any moment. The actual number of ties is free to fluctuate stochastically around this expected value with probabilities determined by the model described below. Having equal activity levels implies that systematic differences in outcome should relate to the pattern of these partnerships and not their magnitude.

For each model, 100 populations were simulated, each for a duration of ten years. The exception is the random model, which served as the basis against which all other populations were compared; the analysis used below requires multiple random populations at each of the infected host sizes observed in the other populations, and obtaining these required 300,000 random populations to be simulated.

The basic framework used in this paper comprises three steps: (1) simulation of dynamic social networks (i.e. in which partnerships form and dissolve over time); (2) simulation of HIV-1 transmission within those networks; and (3) simulation of viral evolution among those infected hosts. The three components of the simulation model were programmed in Delphi, and the project is available for download at http://www.stat.washington.edu/~goodreau.

[1] Note that Strauss and Ikeda refer to the result of this method as a pseudolikelihood; however, in models such as these in which the probability of each tie is independent of the existence of all other ties, their method yields the true likelihood.

Dynamic network simulation: Dynamic networks of social relationships are modeled using the exponential random graph formulation in Eq. 3. The presence of *c* in the denominator makes it difficult to express these probabilities explicitly for all but the smallest of graphs. However, Markov chain Monte Carlo (MCMC) can be used to draw samples from this distribution with the proper frequencies. The algorithm used here is a Metropolis algorithm adopted for network data (Gilks et al. 1996, Crouch and Wasserman 1998), and consists of many loops of the following steps:

1. Randomly select two nodes *i* and *j*.
2. Calculate $\Delta z_{ij}(x)$, the vector of change statistics for all *z* statistics in the model. This refers to the amount by which these statistics change when the relationship between *i* and *j* is toggled from its current state to the opposite state.
3. Calculate the "acceptance probability ratio":

$$L = \frac{\Pr(X_{ij} = toggled \quad value \mid rest \quad of \quad graph)}{\Pr(X_{ij} = current \quad value \mid rest \quad of \quad graph)} = \exp\left(\theta^T \Delta z_{ij}(x)\right).$$

4. If L ≥ 1 (i.e. if the toggled network has an equal or higher probability than the untoggled), accept the toggle as the updated state.
5. If L < 1, then
    a. Select a random number *r* from a uniform (0,1) distribution.
    b. If L > *r* then accept the toggle as the updated state; otherwise retain the original tie value as the updated state.

6. Record the updated state of the entire network.

The basic logic is to select a random movement to a nearby point in the multidimensional network space, use a Metropolis rule to decide whether to accept the step or remain at the current value, and record the position where the chain then rests. The Metropolis rule guarantees that the stationary distribution of the chain equals the probability distribution of Eq. 3 that we were unable to calculate directly (Metropolis et al. 1953). The chain has the basic Markov property that although any two consecutive points in the chain are dependent, as the number of steps between two points in the chain increases this dependence disappears in the limit.

Exponential random graph models with MCMC are generally used to simulate static networks from a probability model of network structure. However, since networks at consecutive steps in the chain are either identical to one another or differ by one tie, this approach also provides a simple way to approximate dynamic networks. In order to make this process dynamic, I allow consecutive steps of the chain to represent consecutive time periods, with each period drawn from an exponential distribution with parameter λ_s. This provides an explicit model of network change while still retaining the instantaneous probability distribution of the static model[2]. All models share the same λ_s, set at the value that would correspond to a mean relationships duration of 1460 days (4 years) in a panmictic population ($\lambda_s = 13.63$).

[2] In general there is no guarantee this approach will yield a chain that resembles a real dynamic network process on the local scale. However, the relatively simple model parameterizations used here should at least ensure that the chain mixes well locally. More realistic methods for dynamic network modeling in the ERGM framework are still in development.

The first chain begins with an empty graph (no ties present) and is run through a one million-iteration burn-in before beginning the clock. This virtually eliminates the dependence on the initial conditions. After the end of one dynamic network, the chain is run though 100,000 steps before beginning the next, in order to prevent dependence among graphs. The outcome of this process is a set of 100 populations for each model (300,000 for the random model), each represented as a vector of all partnerships that existed at some point over the ten year simulation period, including the identity of the actors in the partnership and its starting and ending dates.

Viral transmission: Modeling viral transmission requires a value for infectivity per serodiscordant couple at any given time and a method for bookkeeping those discordant couples. I adopt a constant, universal probability of transmission within each serodiscordant couple. This obviously ignores many sources of heterogeneity, including variation in types of partnership, time since infection of first partner, number of acts engaged in within partnership, number of other partnerships the actors are in, and actor attributes such as STD infection or genetic resistance to HIV-1 infection. However, ignoring these variations makes it possible to state that differences in outcome are due solely to network structure, since that structure is the only thing that varies among the populations. Further work in this vein should benefit from more realistic views of the heterogeneity of infectivity.

Most existing estimates of infectivity are expressed as infectivity per single act or per person-year (i.e. the probability of an active person becoming infected per year), whereas the model I use here requires a measure of infectivity per serodiscordant partnership per

day. I used simulation to derive an infection parameter from information about incidence of HIV-1 infection within American IDU and homosexual male communities, which has been estimated in the range of 1.5-1.9% per year (e.g. Moss et al. 1994, Holmberg et al. 1996). An infectivity level of .0007 infections per serodiscordant partnership per day resulted in an annual incidence rate of 1.57% in the panmictic population and was selected for all simulations. This method is rather *ad hoc*, and is of limited usefulness in generating an estimate with wide applicability.

Since infectivity between serodiscordant partners is assumed to be constant and memoryless, each serodiscordant partnership has an exponentially distributed waiting time until transmission with parameter λ_t. The expected waiting time until the first transmission among all serodiscordant couples is thus also exponentially distributed, with parameter equal to the product of λ_t and the number of serodiscordant partnerships. Of course, this number changes every time a partnership forms or ends and every time a transmission event occurs. Given the list of partnership formation and dissolution times generated in the prior step. I modeled transmission as follows, starting at time t_0:

1. Calculate the number of serodiscordant partnerships s.
2. Draw a random number r from the exponential distribution with parameter $s\lambda_t$.
3. If this number is less than the amount of time remaining until the next change in s, then:
    a. Add r to the current time.
    b. Pick a serodiscordant partnership at random.
    c. Infect the seronegative member of the partnership.

d. Return to Step 1.

4. Otherwise:

    a. Advance to the next time at which s changes.

    b. Return to Step 1.

These steps are repeated until the end of the dynamic network period is reached.

Viral evolution: Each infected host is represented by a single 500-base viral genetic sequence; generating these requires a method for taking account of intrahost viral dynamics without modeling each host's entire quasispecies. I accomplish this by using the intrahost dynamics to define a probability distribution for the timing of sequences' most recent common ancestor, or coalescent. In general an HIV-1-infected host's quasispecies coalesces no earlier than the time at which they were infected, suggesting that people are infected by a single viral particle or that some form of competitive exclusion occurs early in infection. If person A infects person B, then the sequences sampled from A and B must coalesce within A's quasispecies sometime between the moments of A's infection and B's infection. The intrahost dynamics determine the probability distribution for the timing within this interval.

Most of the duration of infection is a long latency period during which the effective population size of the viral pool is relatively constant. Researchers have independently estimated N_e in this period to be on the order of 10^3 (Leigh Brown 1997, Nijhuis et al. 1998 and Rodrigo et al. 1999; but see Rouzine and Coffin 1999). The standard coalescent approximation formula (e.g. Rodrigo and Felsenstein 1999) states that for a

population of constant size, the most recent coalescence event among n sequences in a population of N sequences should come from an exponential distribution with parameter

$$\lambda_c = \frac{n(n-1)}{2N_eG} \qquad \text{[Eq. 4]}$$

where G = generation length. In the case at hand $N_e = 10^3$ and $G = 1.5$ days, an average of the 1.2 day estimate of Rodrigo et al. (1999) and the 1.8 day estimate they cite from personal communication with Perelson. In contrast to this stable equilibrium, the first two months of HIV infection are marked by a rapid expansion and sudden contraction in the effective population size of a host's viral pool. Thus any two sequences within a single host that have not coalesced within 60 days of initial infection can be assumed to coalesce at the point of infection, when rapid expansion begins.

Note that if person *A* infects both person *B* and *C*, then the sampled sequences of *B* and *C* may coalesce with each other (within person *A*) before either of them coalesces with *A*'s sampled sequence. That is to say, two people infected by a common partner may exhibit more similar viral sequences than either do with that partner. Even on a true phylogenetic tree of a local transmission chain, then, two individuals who cluster together may very well have no direct relationship.

As with all coalescent models, reconstruction begins at the ending time of the simulation and proceeds backwards. Let us refer to the series of sequences leading to the sampled sequence for an individual as their *sampling lineage*. The number of potentially coalescing sampling lineages (n) within a given actor's quasispecies changes as one proceeds backwards, decreasing by one each time a coalescent event is determined to

occur, and increasing by one every time a new infection event is reached. The following steps are completed for each infection event (where actor *i* infects actor *j*), beginning with the most recent:

1. Determine the number of potentially coalescing lineages within actor *i* (n_i), including *i*'s sampling lineage, *j*'s sampling lineage, and the sampling lineage of anyone else *i* has infected more recently and whose sampling lineage has not yet coalesced with that of *i*.
2. Determine the most recent coalescence time using Eq. 4.
    a. If this time is more recent than any other infection event involving the actor and is not within 60 days of the actor's own infection date, then a coalescence event occurs. If there are more than two potentially coalescing lineages in the actor, the two that coalesce are selected randomly. If $n_i > 1$ still then return to Step One.
    b. If the time is less recent than another infection event involving the actor, no coalescence occurs yet.
    c. If the time is less recent than the actor's infection date or less than 60 days more recent than the actor's infection date, all sequences in the actor coalesce at her infection date.

In simulating HIV-1 mutation within these lineages, I use a model that incorporates much of the known heterogeneity in HIV-1 microevolution. This includes Leitner and Albert's (1999) mean mutation rate for the *env* gene, 6.7×10^{-3} substitutions per base per

year, or 1.8 x 10^{-5} substitutions per base per day. It also includes the *env* nucleotide frequencies and transition matrix from Leitner et al. (1997), both shown in Table 3, and gamma-distributed inter-site mutation rates, using their point estimate $\alpha = 0.384$ for *env*. The same set of site-specific mutation rates drawn from this distribution was used in each simulation. The instantaneous mutation rate for site x (μ_x) is thus the product of the overall mutation rate, the relative mutability of the x's current nucleotide (taken from the main diagonal of Table 3b), and x's gamma mutation factor. Note that the expected value of both the second and third factor is 1, so that the average mutation rate is unaffected. The effects of selection and recombination are ignored.

Simulation of mutation begins at time t_0, when the first person in the population is infected from an outside source, and proceeds forward in time. The initial sequence for this person is generated randomly, with nucleotide probabilities chosen according to the frequencies given in Table 3b. Mutations then occur along this lineage by repeatedly calculating the current mutation rate μ_x for each site x; drawing a time until next mutation t_x for each site from an exponential distribution with mean equal to μ_x; and selecting $\min\{t_1 \ldots t_{500}\}$. The new nucleotide at the mutating site is selected with the relative probabilities from the corresponding row in Table 3b. This process is repeated for all other infected actors, with the initial sequence in each's sampling lineage matching that of the sequence to which it coalesces.

RESULTS

In looking at the differences in genetic outcomes between the models, it is important to realize that these could simply result from different population sizes, i.e. if there are systematically more actors infected in one type of population than another. In order to examine differences in genetic diversity net of potential differences in population size, the first step is thus to test whether such differences exist. Figure 1 shows the distribution of the number of infected individuals across the 100 runs for each of the six structured mixing models and 300,000 for the random model. Summary statistics for these distributions are contained in Table 4, including the results of a test of Kullback-Leibler distances between each distribution and that of the random population. P-values were obtained by sampling 100 runs from the random pool, calculating the Kullback-Leibler distance for this sample from the original distribution, and repeating 1000 times. These scores show that subdivision into two groups has little effect on prevalence, even when those groups are strongly isolated. All other models have prevalence distributions that were significantly different from the random model, with the clustering models having a lower average prevalence and the core and bridge models having higher prevalence. These results are consistent with the previous work described above showing the effects of network structure on prevalence. It is important to remember that these simulations were of finite duration in a closed population, so that these prevalence levels do not represent any kind of long-term endemic prevalence figure.

Given these systematic differences in census population size, I compared observed mismatch means from the various structured populations directly to those observed for

panmictic populations that yielded the same number of infected hosts. This comparison required a sufficient number of runs of the random model at each of the infected population sizes observed in the other models (i.e. from one infected host up through 133). The random model rarely yielded host sizes at the upper end of this range; hence the need for 300,000 runs. I calculated the mean of the mismatch distribution δ for each of these 300,000 runs. The distribution of δ for each host population size is summarized in Figure 2 by the mean and the upper and lower 95% quantiles. (Note that the distribution here refers to the distribution of δ across the many runs of a given size, not to the distribution of pairwise differences directly). Medians are not plotted since the distribution of δ was roughly normal at each size, making the median virtually indistinguishable from the mean on this graph. It is clear from Figure 2 that the regularity in these distributions breaks down around a host size of 130, when sample sizes become small (<25 runs).

The bounds containing 95% of the observations at each size allow us to see where the outcomes from the structured models fall relative to the greatest part of the panmictic populations. Outcomes for these other populations are graphed against the panmictic populations in Figure 3. We see that many have significantly greater genetic diversity than that found in comparably sized randomly mixing host populations, while only one run among all of the models has significantly less. Table 5 shows the number of runs outsides the bounds of the random populations for each model.; this makes it clear that for all but the bridges model, effective population size lies outside the range of stochastic variation of the random population a considerable fraction of the time. These differences thus seem systematic rather than stochastic; given this fact, the last column informs us of

the average fraction by which the genetic diversity (and therefore the effective population size, since they are proportional according to Eq. 1) of each model lies above the mean in a panmictic population of the same size. This value equals:

$$\frac{1}{100}\sum_{i=1}^{100}\frac{\delta_{i,x}-\overline{\delta}_{x,n_{i,rand}}}{\overline{\delta}_{x,n_{i,rand}}} \qquad \text{[Eq. 5]}$$

where $\delta_{i,x}$ represents the mean of the mismatch distribution for run i of mixing model x, $n_{i,x}$ represents the number of infected hosts in run i of mixing model x, and $\overline{\delta}_{x,n}$ is the mean of δ for all runs of model x that that have n infected hosts. This value ranges from a low of 10% for the bridges model to a high of 24% for the two strongly assortative models, with a mean of 18% across all six models. The populations that have greater genetic diversity than the random population generally see viral spreads rapidly in one section of the population; the virus may then take a while to spread into another cluster, or may not at all. This leads to a number of early divergence events (because of the rapid initial spread) but a lower population size than might be expected if that rate had continued.

Since host population size in the early stages of an epidemic often grows roughly exponentially, it is interesting to compare the observed values of δ to those generated by Eq. 2 for the appropriate value of N_0. Of course this equation assumes that mixing among all viruses in the metapopulation is random, which is obviously not true; even if hosts chose their sexual partners randomly, the viral pools of different infected hosts are not able to mix together. The process of sampling exactly one sequence from each host adds

a further wrinkle. Such a comparison, then, allows us to see the combined effect of departures from pure exponential growth in host size, isolation of the hosts' viral pools from each other, and the sampling constraint of one-sequence per host, but does not allow us to separate these effects out from one another.

The definition of exponential growth states that at time t in the past, $N_t = N_0 e^{-rt}$. Since we know that $N_t = 1$ at $t = 2433.3$ generations in the past (3650 days / 1.5 days per generation = 2433.3), we can specify $r = \ln(N_0)/2433.3$. This reduces Eq. 2 to

$$\delta = 2*2.7*10^{-5} * \frac{2433*\left(\ln\left(\frac{N_0 \ln(N_0)}{2433}\right) - \gamma\right)}{\ln(N_0)} \qquad \text{[Eq. 6]}$$

with N_0 as the only unknown. (2.7×10^{-5} is the mutation rate per site per generation, equal to the daily mutation rate times 1.5 days per generation). Since each host infected more than 60 days contains an effective viral pool of 10^3, I approximate the total viral pool as 10^3 times the number of infected hosts. With these assumptions, the δ predicted for each infected host size is shown in Figure 4, and the ratio of this value to that observed in the random model (i.e. complete panmixis vs. viral metapopulation with panmictic hosts) is shown in Figure 5. The distribution of δ across different population sizes follows a qualitatively similar pattern in the two, with the combined effects of metapopulation structure and departures from exponential growth reducing this value in the simulated population by a factor of 2.0-2.6.

DISCUSSION

HIV-1 research is unique for the large role played by population genetics in the practical applications of the field; accurate estimates of viral population size and dynamics, both within and between hosts, are important for applications as wide-ranging as reconstructing general patterns of viral spread, understanding differences in transmissibility of viral genotypes, and testing hypotheses for the transmission of antiviral resistance. The work here was a first step in understanding how the network of behaviors that spread HIV-1 in the first place can affect the relationship between census size and effective population size at the community level. This relationship differed by an average of 18% over that found in randomly mixing populations with the same host population size, a figure that could easily be greater in populations with more complicated and realistic mixing patterns.

In this study I have purposefully left out many forms of complexity that will need to be considered by future work in this vein. Chief among these are the birth/death dynamics of real populations and heterogeneity in infectivity by partnership type and duration. The expected effect of ignoring the latter form of heterogeneity is to underestimate the importance of short relationships as sources of viral spread. This will have its greatest effect in populations in which short partnerships are relatively important, such as those involving commercial sex workers. In the real world, partnership type and duration may also have relevant covariates such as frequency of condom use. Future work might consider classifying partnerships into a small number of categories based on duration and actor attributes, with different data-derived infectivities per category.

This work clearly demonstrated the effect that host mixing structure may have on the population genetics of HIV-1; however, this was done through the use of simulation rather than analytically, and used only mismatch distributions rather than full phylogenetic information. Ideally we would like a general method for incorporating mixing structure directly into formulas for expected values of genetic parameters such as the shape of the mismatch distribution or the number of lineages through time in a reconstructed tree. Hopefully, the exponential random graph model framework will eventually provide a means for doing so. This approach possesses the generality to describe a far greater range of mixing rules than those examined here, including patterns in which individual partnership probabilities depend directly on the status of other partnerships. For many of these cases, however, the interpretability of the resulting parameters is not straightforward, although progress is now being made on this front. As this advances, we will hopefully gain the ability to incorporate realistic models of human sexual behavior into our analysis of the population genetics of pathogens that are spread as a result of this behavior.

The practical importance of the observed effects of network structure will of course depend on the questions one is examining when applying such methods. In any type of study that uses community-level data on HIV-1 genetic diversity to test hypotheses of HIV-1 evolution, the signature of interest could be masked by the confounding effects of metapopulation dynamics. One example is the use of phylogenetic methods to explore questions of viral selection pressures between hosts such as those resulting from anti-viral drug resistance. Determining the potential for the spread of drug resistance mutations is of fundamental public health importance, and mathematical models have begun to appear

that address this question (Blower et al. 2001). Much remains to be done, however, and phylogenetic studies could contribute greatly to this important enterprise. When examining community-level viral sequence data for signatures of selection, we will have greater power if we have clear and accurate ideas of the genetic patterns that should appear in real populations in the absence of selection.

The author would like to thank Martina Morris, Ken Weiss, Mark Handcock, Jim Wood, Andy Clark, Anne Buchanan, Steve Koester, Anna Baron, John Potterat and Jamie Jones. This research was funded by graduate fellowships from the National Science Foundation and the Population Council and by a National Institutes of Health STD/AIDS Research Training Grant to the University of Washington Center for AIDS and STD (5T32AI007140-25).

LITERATURE CITED

AHMAD, N., B. M. BAROUDY, R. C. BAKER and C. CHAPPEY, 1995 Genetic analysis of human immunodeficiency virus type 1 envelope V3 region isolates from mothers and infants after perinatal transmission. J. Virol. **69:** 1001-12.

BESAG, J., 1974 Spatial interaction and the statistical analysis of lattice systems. J. Royal Stat. Soc. Ser. B **36:** 192-236.

BLOWER, S. M., A. N. ASCHENBACH, H. B. GERSHENGORN and J. O. KAHN, 2001 Predicting the unpredictable: transmission of drug-resistant HIV. Nat. Med. **7:** 1016-20.

CROUCH, B., and S. WASSERMAN, 1998 Fitting p*: Monte Carlo Maximum Likelihood Estimation. *Sunbelt XVIII and Fifth European International Social Networks Conference*, Sitges, Spain.

FELSENSTEIN, J., 1992 Estimating effective population size from samples of sequences: inefficiency of pairwise and segregating sites as compared to phylogenetic estimates. Genet. Res. **59:** 139-47.

FRANK, O., and D. STRAUSS, 1986 Markov Graphs. JASA **81:** 832-842.

FRIEDMAN, S. R., A. NEAIGUS, B. JOSE, R. CURTIS, M. GOLDSTEIN et al., 1997 Sociometric risk networks and risk for HIV infection. Am. J. Public Health **87:** 1289-1296.

GARNETT, G. P., J. P. HUGHES, R. M. ANDERSON, B. P. STONER, S. O. ARAL et al., 1996 Sexual mixing patterns of patients attending sexually transmitted diseases clinics. Sexually Trans. Dis. **23:** 249-257.

GILKS, W. R., S. RICHARDSON and D. J. SPIEGELHALTER, 1996 *Markov chain Monte Carlo in practice*. Chapman & Hall, London.

GRASSLY, N. C., P. H. HARVEY and E. C. HOLMES, 1999 Population dynamics of HIV-1 inferred from gene sequences. Genetics **151:** 427-438.

HANSKI , I. A., and M. A. GILPIN, 1997 *Metapopulation biology: ecology, genetics, and evolution*. Academic Press, San Diego.

HARALDSDOTTIR, S., S. GUPTA and R. M. ANDERSON, 1992 Preliminary studies of sexual networks in a male homosexual community in Iceland. J. AIDS **5:** 374-381.

HOLMBERG, S. D., 1996 The estimated prevalence and incidence of HIV in 96 large US metropolitan areas. Am. J. Public Health **86:** 642-54.

HOLMES, E. C., O. G. PYBUS and P. H. HARVEY, 1999 The molecular population dynamics of HIV-1, pp. 177-207 in *The Evolution of HIV*, edited by K. A. CRANDALL. Johns Hopkins University Press, Baltimore.

KINGMAN, J. F. C., 1982 On the genealogy of large populations. J. Appl. Prob. **19A**: 27-43.

KLOVDAHL, A. S., J. J. POTTERAT, D. E. WOODHOUSE, J. B. MUTH, S. Q. MUTH et al., 1994 Social networks and infectious disease: the Colorado Springs study. Soc. Sci. Med. **38:** 79-88.

LAUMANN, E. O., J. H. GAGNON, R. T. MICHAEL and S. MICHAELS 1994 *The social organization of sexuality : sexual practices in the United States*. University of Chicago Press, Chicago.

LEIGH BROWN, A. J., 1997 Analysis of HIV-1 env gene sequences reveals evidence for a low effective number in the viral population. Proc. Natl. Acad. Sci. U S A **94:** 1862-5.

LEIGH BROWN, A. J., D. LOBIDEL, C. M. WADE, S. REBUS, A. N. PHILLIPS et al., 1997 The molecular epidemiology of human immunodeficiency virus type 1 in six cities in Britain and Ireland. Virology **235:** 166-177.

LEITNER, T., and J. ALBERT, 1999 The molecular clock of HIV-1 unveiled through analysis of a known transmission history. Proc. Natl. Acad. Sci. USA **96:** 10752-10757.

LEITNER, T., D. ESCANILLA, C. FRANZEN, M. UHLEN and J. ALBERT, 1996 Accurate reconstruction of a known HIV-1 transmission history by phylogenetic tree analysis. Proc. Natl. Acad. Sci. U S A **93:** 10864-9.

LEITNER, T., S. KUMAR and J. ALBERT, 1997 Tempo and mode of nucleotide substitutions in gag and env gene fragments in human immunodeficiency virus type 1 populations with a known transmission history. J. Virol. **71:** 4761-70.

MARTIN, J. L., 1987 The impact of AIDS on gay male sexual behavior patterns in New York City. Am. J. Pub. Health **77:** 578-581.

MCCUTCHAN, F. E., B. L. UNGAR, P. HEGERICH, C. R. ROBERTS, A. K. FOWLER et al., 1992 Genetic analysis of HIV-1 isolates from Zambia and an expanded phylogenetic tree for HIV-1. J. AIDS **5:** 441-9.

METROPOLIS, N., A. W. ROSENBLUTH, M. N. ROSENBLUTH, A. H. TELLER and E. TELLER, 1953 Equation of state calculations by fast computing machines. J. Chem. Phys. **21:** 1087-1092.

MORRIS, M., 1991 A log-linear modeling framework for selective mixing. Math. Biosci. **107:** 349-77.

MORRIS, M., and L. DEAN, 1994 Effect of sexual behavior change on long-term human immunodeficiency virus prevalence among homosexual men. Am. J. Epidemiol. **140:** 217-32.

MORRIS, M., and M. KRETZSCHMAR, 1997 Concurrent partnerships and the spread of HIV. AIDS **11:** 641-648.

MORRIS, M., C. PODHISITA, M. J. WAWER and M. S. HANDCOCK, 1996 Bridge populations in the spread of HIV/AIDS in Thailand. AIDS **10:** 1265-1271.

MOSS, A. R., K. VRANIZAN, R. GORTER, P. BACCHETTI, J. WATTERS et al., 1994 HIV seroconversion in intravenous drug users in San Francisco, 1985- 1990. AIDS **8:** 223-31.

NEE, S., R. M. MAY and P. H. HARVEY, 1994 The reconstructed evolutionary process. Philos. Trans. R. Soc. Lond. **344:** 305-11.

NEE, S., E. C. HOLMES, A. RAMBAUT and P. H. HARVEY, 1995 Inferring population history from molecular phylogenies. Philos. Trans. R. Soc. Lond. **349:** 25-31.

NIJHUIS, M., C. A. BOUCHER, P. SCHIPPER, T. LEITNER, R. SCHUURMAN et al., 1998 Stochastic processes strongly influence HIV-1 evolution during suboptimal protease-inhibitor therapy. Proc. Natl. Acad. Sci. U S A **95:** 14441-6.

ORUBULOYE, I. O., J. C. CALDWELL and P. CALDWELL, 1991 Sexual networking in the Ekiti District of Nigeria. Studies in Family Planning **22:** 61-73.

OU, C. Y., C. A. CIESIELSKI, G. MYERS, C. I. BANDEA, C. C. LUO et al., 1992 Molecular epidemiology of HIV transmission in a dental practice. Science **256:** 1165-71.

PYBUS, O. G., A. RAMBAUT and P. H. HARVEY, 2000 An integrated framework for the inference of viral population history from reconstructed genealogies. Genetics **155:** 1429-37.

RODRIGO, A. G., and J. FELSENSTEIN, 1999 Coalescent approaches to HIV population genetics, pp. 233-272 in *The Evolution of HIV*, edited by K. A. CRANDALL. Johns Hopkins University Press, Baltimore.

RODRIGO, A. G., E. G. SHPAER, E. L. DELWART, A. K. IVERSEN, M. V. GALLO et al., 1999 Coalescent estimates of HIV-1 generation time in vivo. Proc. Natl. Acad. Sci. U S A **96:** 2187-91.

ROTHENBERG, R. B., C. STERK, K. E. TOOMEY, J. J. POTTERAT, D. JOHNSON et al., 1998 Using social network and ethnographic tools to evaluate syphilis transmission. Sexually Trans. Dis. **25:** 154-160.

ROUZINE, I. M., and J. M. COFFIN, 1999 Linkage disequilibrium test implies a large effective population number for HIV in vivo [see comments]. Proc. Natl. Acad. Sci. U S A **96:** 10758-63.

SLATKIN, M., and R. R. HUDSON, 1991 Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. Genetics **129:** 555-62.

STRAUSS, D., and M. IKEDA, 1990 Pseudolikelihood estimation for social networks. J. Am. Stat. Soc. **85:** 204-212.

VARTANIAN, J. P., A. MEYERHANS, M. HENRY and S. WAIN-HOBSON, 1992 High-resolution structure of an HIV-1 quasispecies: identification of novel coding sequences. AIDS **6:** 1095-8.

WASSERMAN, S., and P. PATTISON, 1996 Logit models and logistic regression for social networs: I. An introduction to Markov graphs and p*. Psychometrika **60:** 401-426.

WATTERSON, G. A., 1975 On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol. **7:** 256-276.

WATTS, C. H., and R. M. MAY, 1992 The influence of concurrent partnerships on the dynamics of HIV/AIDS. Math. Biosci. **108:** 89-104.

Table 1: Population mixing models

Model	Abbr.	Description
Random	Rand	All partnerships equally likely
Two subgroups, weak assortivity	2-strong	Population divided into two groups of 100; 150 intragroup ties and 50 intergroup ties expected
Two subgroups, strong assortivity	2-weak	Population divided into two groups of 100; 195 intragroup ties and 5 intergroup ties expected
Eight subgroups, weak assortivity	8-strong	Population divided into eight groups of 25; 150 intragroup ties and 50 intergroup ties expected
Eight subgroups, weak assortivity	8-weak	Population divided into eight groups of 25; 195 intragroup ties and 5 intergroup ties expected
Core/periphery	Core	Population divided into active group of 25 and periphery of 175; 30 intra-core ties expected (10 times more that expected by chance)
Bridges	Bridges	Population divided into 95 husbands, 95 wives, and 10 non-married females. Spouses remain married throughout, 105 ties between husbands and other females expected

In all of the assortative populations (2-strong, 2-weak, 8-strong, 8-weak), individuals chose partners preferentially from within their own cluster, but all clusters had equal activity. In the core/periphery group, core members were more active but did not choose other core partners with any greater probability than they did periphery members, once controlling for overall activity levels.

Table 2: Graph statistics and their associated parameters for each mixing model

Model	Z	θ
Random	# of ties	-4.590
2-strong	# of ties	-3.907
	# of inter-group ties	-3.693
2-weak	# of ties	-4.174
	# of inter-group ties	-1.119
8-strong	# of ties	-2.426
	# of inter-group ties	-4.895
8-weak	# of ties	-2.708
	# of inter-group ties	-2.387
Core	# of ties	-4.739
	# of intra-core ties	2.542
Bridges	# of male-single female ties	-2.085
	# of male-male ties	$-\infty$
	# of female-female ties	$-\infty$

Table 3: Allele frequencies and transition matrix for *env*

A) allele frequencies

A	.4627
C	.1474
G	.1598
T	.2302

B) transition matrix

	A	C	G	T
A	-0.8351	0.2519	0.4599	0.1233
C	0.7909	-1.3932	0.0574	0.5449
G	1.3318	0.0529	-1.5406	0.1558
T	0.2477	0.3488	0.1081	-0.7047

(data from Leitner et al. 1997)

Table 4: Distribution of infected hosts among 300,000 runs of the random (panmictic) model and 100 runs of each other model

Model	**Mean**	**St. dev.**	**KL distance**	**p**
Random	29.8	26.1	---	---
2-strong	31.1	22.4	0.240	.56
2-weak	26.1	22.1	0.225	.69
8-strong	14.5	8.0	1.215	< .001
8-weak	22.8	20.1	0.380	.01
Core	65.9	33.4	0.718	< .001
Bridges	53.8	44.5	1.159	< .001

Kullback-Leibler distances and *p*-values are for a test comparing the given distribution to the random model distribution.

Table 5: Mismatch means compared to the random model

Model	**Total above random 97.5% quantile**	**Total below random 2.5% quantile**	**Average fractional difference in effective population size from random mixing**[1]
2-strong	32	0	0.237
2-weak	20	0	0.168
8-strong	19	0	0.238
8-weak	24	0	0.219
Core	22	0	0.133
Bridges	5	1	0.104

[1]See text for formula

Figure 1: Distribution of infected population resulting from 300,000 runs of the random (panmictic) population and 100 runs of each of the structured models

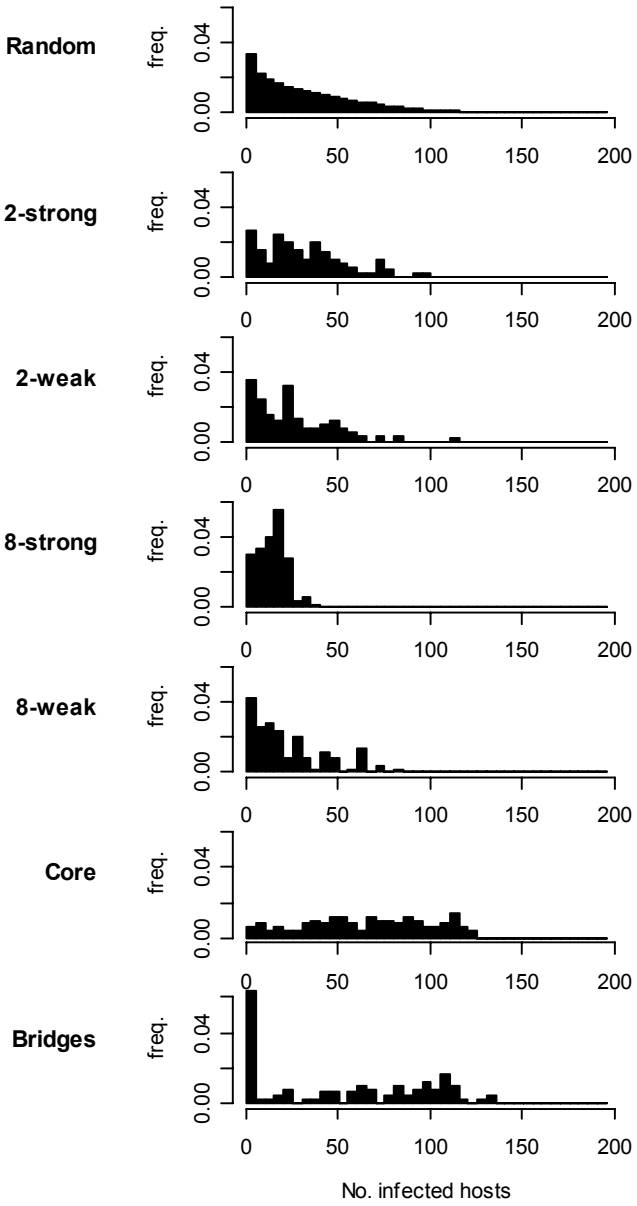
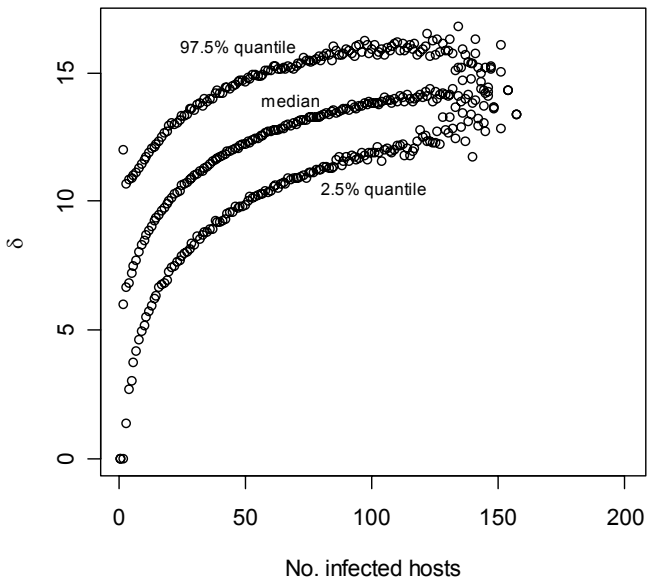


Figure 2: Mismatch mean for 300,000 runs of the random model, by infected population size



Mean values are indistinguishable from medians. Values above $x = 130$ show the effects of small sample sizes.

Figure 3: Structured models compared to random model for mean mismatch by population size

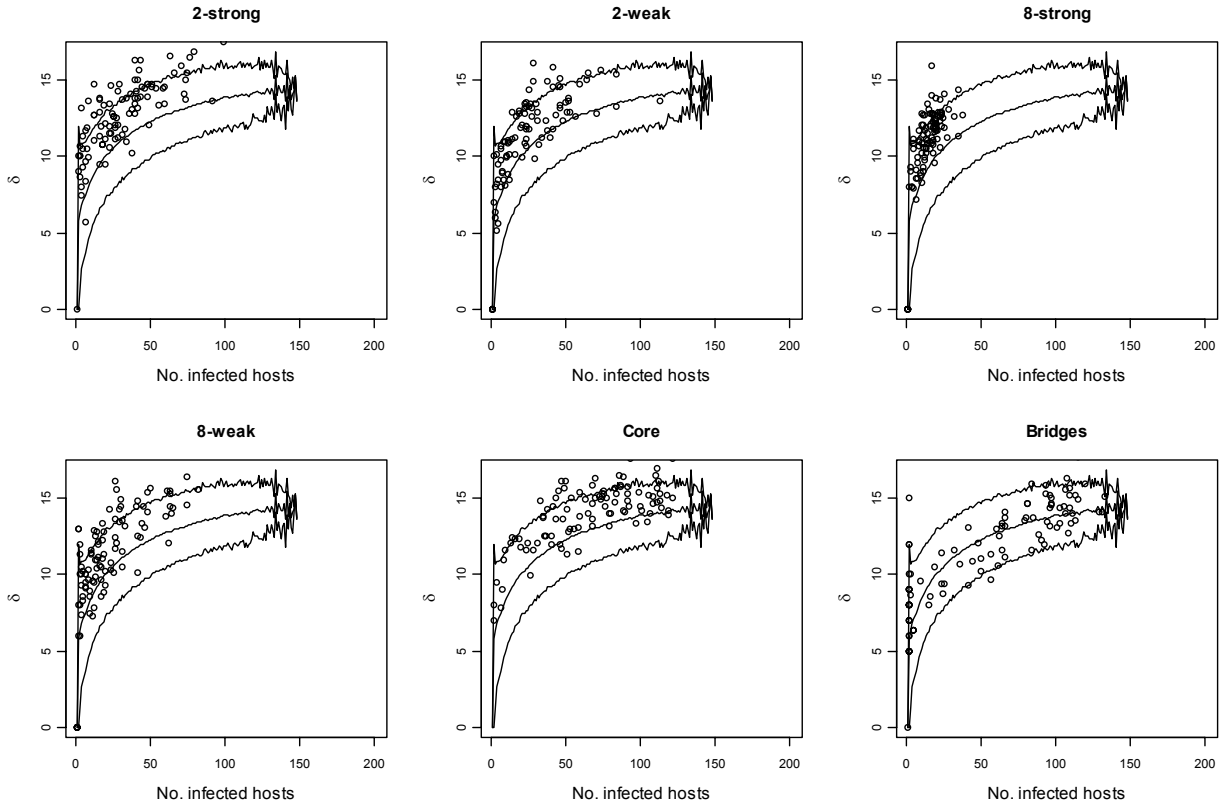


Figure 4: Values of δ expected for a completely panmictic viral population equivalent to observed host sizes

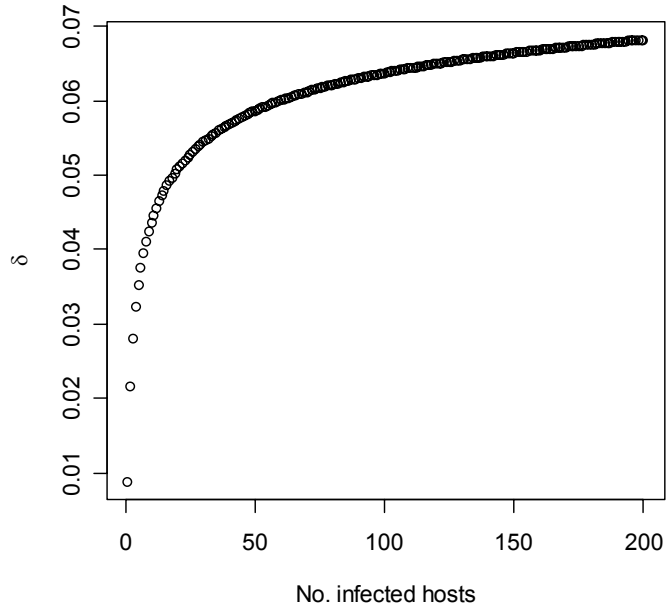


Figure 5: Ratio of observed values of δ to those expected in a completely panmictic viral population of the same size