# Effects of multiple sources of genetic drift on pathogen variation within hosts

David A. Kennedy[1,2,3] and Greg Dwyer[2,4]

[1]Center for Infectious Disease Dynamics, Pennsylvania State University
[2]Ecology and Evolution, University of Chicago
[3]dak30@psu.edu.
[4]To whom correspondence should be addressed: gdwyer@uchicago.edu.

**Running title**: Multiscale drift affects pathogen variation

**Keywords**: host-pathogen ecology, genetic drift, nested model, sequence data, within-host diversity, phylogenetic inference, population bottleneck, demographic stochasticity

**Words in abstract**: 146
**Words in main text**: 4469
**Number of references**: 57
**Number of figures**: 5

**Statement of authorship**: DK and GD conceived of the study, constructed the models, analyzed the data, and wrote the manuscript
**Statement of authorship**: DK and GD conceived of the study, constructed the models, analyzed the data, and wrote the manuscript
**Data and Code Availability**: All sequence data generated in this study are available through the Sequence Read Archive (SRA) of the National Center for Biotechnology Information (NCBI) under BioProject ID PRJNA386565. All author-generated code used in this study is available from our GitHub repository: `https://github.com/dkenned1/KennedyDwyer`.

**Changes in pathogen genetic variation within hosts alter the severity and spread of infectious diseases, with important implications for clinical disease and public health. Genetic drift may play a strong role in shaping pathogen variation, but host-pathogen models that describe the effects of drift typically focus on dynamics only within hosts or only between hosts, while models that incorporate drift at both scales oversimplify pathogen ecology in other ways. We used whole-genome sequencing to quantify variation within 143 field-collected gypsy moth larvae infected with a baculovirus and we compared the ability of different host-pathogen models to explain differences in variation across hosts. A model in which drift acts during transmission bottlenecks and during pathogen growth within hosts accurately reproduces the data, whereas models that neglect or simplify drift fail to explain the data. Our results illustrate the usefulness of mechanistic ecological models in explaining pathogen evolution.**

## Introduction

Pathogen genetic variation can have important consequences for human health, in both clinical and epidemiological settings (Alizon et al. 2011). In particular, high variation within hosts can lead to severe disease symptoms within individuals, and rapid disease transmission within populations (Read and Taylor 2001; Vignuzzi et al. 2006). An understanding of the mechanisms determining pathogen variation might therefore lead to novel interventions, reducing the toll of infectious diseases. Development of such an understanding requires quantification of the effects of population processes on pathogen genetic variation.

Models relating population processes to pathogen genetic variation, however, often greatly simplify pathogen biology. Selection-mutation models, for example, often assume that

pathogen populations are effectively infinite (Lorenzo-Redondo et al. 2016). Models that allow for finite pathogen population sizes typically neglect pathogen population processes either within hosts in acute infections (Koelle et al. 2006) or between hosts in chronic infections (Pennings et al. 2014). Models that attempt to capture both of these scales of disease dynamics have assumed either that pathogen population growth within hosts is very simple, with no effects of the immune system (Klinkenberg et al. 2017; Leonard et al. 2017), or that pathogen bottlenecks at transmission are complete (Ypma et al. 2013; Didelot et al. 2014; Klinkenberg et al. 2017), so that every infection begins as a clonal lineage. These simplifications might strongly alter conclusions about the effects of genetic drift on pathogen diversity, and indeed have been highlighted as some of the key challenges of phylodynamic inference (Frost et al. 2015).

Genetic drift is a change in an allele's frequency due to the chance events that befall individuals. The effects of drift are thus strongest in small populations, in which a few events can have a large impact (Nagylaki 1992). The high population sizes typical of severe infections have led some authors to argue that drift has little effect on pathogens (Kouyos et al. 2006; Maldarelli et al. 2013), but pathogen population sizes typically fluctuate by several orders of magnitude over the course of an infection. It therefore seems likely that, when pathogen populations are small and variable, pathogen genetic variation will be strongly affected by drift. Indeed, analyses that allow for finite population sizes have shown that drift has at least weak effects on some pathogens (Pennings et al. 2014).

New infections are typically initiated by small pathogen population sizes within hosts (Gutiérrez et al. 2012), leading to bottlenecks at the time of transmission that may drive genetic drift. Pathogen population sizes within hosts can also remain small for substantial periods following exposure (Kennedy et al. 2014). In small populations, chance events such as the timing of reproduction can strongly influence population growth (Kennedy et al. 2014), an

effect known as “demographic stochasticity” (Kot 2001). When the effects of demographic stochasticity are strong, chance may allow some virus strains to replicate and survive while others go extinct, providing a second source of genetic drift that we refer to as “replicative drift”. Note that we use the term “strain” here and throughout to mean a population of virus particles that are identical to each other in their genetic sequence.

Many previous studies of genetic drift in pathogens have focused only on population processes that operate within hosts, either during experiments with model organisms, or during the treatment of human patients (Abel et al. 2015). Pathogen variation in nature, however, is also affected by processes that operate at the host population level, such as fluctuating infection rates during epidemics (Grenfell et al. 2004). Studies of Ebola (Azarian et al. 2015) and tuberculosis (Lee et al. 2015), for example, have shown that variation at the population level cannot be explained by natural selection, which means that variation must instead be due to neutral processes that presumably include genetic drift.

Genetic drift in pathogens may thus be driven by population processes at multiple scales. These multiple scales can be incorporated into a single framework by constructing “nested” models, in which sub-models of within-host pathogen population growth are nested in models of between-host pathogen transmission (Mideo et al. 2008). The computing resources necessary to analyze such complex models have only become available recently, however, and so it has not been clear whether sufficient data exist to test nested models of drift (Gog et al. 2015). Indeed, even for models that assume that population size effects are constant across hosts, robust tests of the model predictions require both genetic data and mechanistic epidemiological models (Didelot et al. 2014), a combination that is rarely available. Whether nested models can be of practical use for understanding pathogen genetic variation in nature is therefore unclear.

For baculovirus diseases of insects, pathogen population processes have been intensively studied at both the host population level (Elderd 2013), and at the individual host level (Kennedy

et al. 2014). Baculoviruses cause severe epizootics (= epidemics in animals) in many insects (Moreau and Lucarotti 2007), including economically important pest species such as the gypsy moth (*Lymantria dispar*) that we study here (Woods and Elkinton 1987). Collection and rearing protocols for the gypsy moth have long been standardized (Elkinton and Liebhold 1990), and so previous studies of the gypsy moth baculovirus have produced statistically robust estimates of the parameters of both within-host (Kennedy et al. 2014) and between-host (Fuller et al. 2012; Elderd et al. 2008) models. Moreover, collection of large numbers of virus-infected individuals is straightforward (Woods and Elkinton 1987), making it possible to use high-throughput sequencing methods to characterize pathogen diversity across many virus-infected hosts. Here we therefore use a combination of whole-genome sequence data and parameterized, nested models to quantify the effects of genetic drift on the gypsy moth baculovirus. We show that a mechanistic model of genetic drift can explain variation in this pathogen, but only if the model takes into account the effects of drift at multiple scales of biological organization.

Our work differs from previous efforts to understand genetic variation in pathogens in that we explicitly allow for within-host population dynamics, including nonlinear effects of the host immune system. Previous work has attempted to infer rates of disease transmission from "phylodynamic" models, which infer disease demography and pathogen evolution from genetic data (Grenfell et al. 2004). Here we instead begin with an existing population process model that has already been fit to epidemiological data, and we use it to predict pathogen genetic data. We are thus testing the extent to which disease demography can be used to predict neutral pathogen evolution.

## Materials and Methods

Like many insect baculoviruses, the gypsy moth baculovirus is transmitted when larvae accidentally consume infectious particles known as "occlusion bodies" while feeding on foliage

(Elderd et al. 2008). If the virus population resulting from the consumption of occlusion bodies reaches a sufficiently large size during within-host growth, the larva dies, releasing new occlusion bodies onto the foliage. Occlusion bodies on foliage are then available to be consumed by additional conspecifics (the virus is species specific (Moreau and Lucarotti 2007)), leading to very high infection rates in high density populations (Woods and Elkinton 1987). Genetic drift in the gypsy moth baculovirus can therefore be affected by population processes at multiple scales, including within individual hosts and across the host population.

As is often the case for pathogens, exposure to baculovirus results in an initial population of only a few occlusion bodies (Zwart et al. 2009), and the population size in the host remains small for a substantial period of time following exposure (Kennedy et al. 2014). Our models of pathogen growth within hosts therefore track population sizes from the initial population bottleneck through the stochastic growth of the pathogen population, until death or recovery. Our models thus explicitly include genetic drift (fig. 1).

Our within-host models are based on birth-death models (Kot 2001), which describe probabilistic changes in population sizes over time. In birth-death models, the probability of a birth or a death in a small period of time increases with the population size (Renshaw 1991). When the population size is small in a birth-death model, it is possible for extinction to occur due to a chance preponderance of deaths over births, even if the per-capita birth rate exceeds the per-capita death rate. Birth-death models are thus well suited to describe the demographic stochasticity that underlies replicative drift.

In our within-host birth-death model, pathogen extinction is equivalent to the clearance of the infection by the host. If the pathogen does not go extinct, its population eventually becomes large enough that the effects of stochasticity are negligible, leading to host illness when the population reaches an upper threshold (Saaty 1961), or death in the case of baculoviruses. In traditional birth-death models, which include only linear effects of population size (Kot

2001), each virus particle has a constant probability of reproducing or dying. In previous work, however, we showed that linear models are insufficient to explain data on the speed of kill of the gypsy moth baculovirus, whereas models that allow for nonlinearities due to the immune system produce a better explanation for the data (Kennedy et al. 2014).

Our within-host model thus describes the outcome of a process that begins with the insect's immune system releasing chemicals that active the phenol-oxidase pathway. This release causes virus particles to be encapsulated and destroyed by host immune cells, and it also incapacitates the immune cell (Ashida and Brey 1998; Trudeau et al. 2001). Our models thus follow standard predator-prey-type immune-system models (Alizon and van Baalen 2008), in which the pathogen is the prey, and the immune cells are the predator, except that here the immune cells do not reproduce over the timescale of a single infection. The pathogen population in the model may then be driven to zero because of interactions with the host immune system, or it may persist long enough to overwhelm the host immune system, leading to exponential pathogen growth and eventual host death. The particular outcome that occurs depends on the initial pathogen population size and on demographic stochasticity during the infection.

The initial pathogen population size within a host follows a Poisson distribution (fig. 1 (Kennedy et al. 2014)). If the infecting cadaver is comprised of multiple strains, the model draws an initial population size for each of the strains from a multinomially distributed random variate, such that the probability of sampling a particular strain from the infecting cadaver depends on the frequency of that strain in the cadaver. This process creates a transmission bottleneck. Next, the model tracks the population size of each virus strain over the course of the infection. Changes in the relative frequencies of these strains over time creates replicative drift. The host dies when the total pathogen population size from an exposure exceeds an upper threshold. The frequency of virus strains at the time of host death then determines the frequency of strains in the newly generated cadaver.

We model pathogen dynamics at the scale of the entire host population first by using a stochastic Susceptible-Exposed-Infected-Recovered or "SEIR" model to describe epizootics (in our case the infected $I$ class consists of infectious cadavers in the environment, which we symbolize as $P$ for the pathogen). Our SEIR model is modified to allow hosts to vary in infection risk, an important feature of gypsy moth virus transmission (Dwyer et al. 1997; Elderd et al. 2008), and to allow exposed hosts to be re-exposed, because infected gypsy moth larvae continue to consume foliage until shortly before death (Eakin et al. 2014). For computational convenience (Wearing et al. 2005), most SEIR models assume that the times between hosts becoming exposed and becoming infectious follow a gamma distribution (Keeling and Rohani 2008). We instead allow the speed of kill to be determined by our within-host model, so that the within-host model is nested inside the stochastic SEIR model. As in the within-host model, the frequency of different virus strains at the population scale can drift due to chance events, such as the exposure of hosts to one cadaver and not another. Our between-host model therefore adds an additional source of drift to our nested models, in addition to the transmission bottlenecks and replicative drift that occur in the within-host model.

Over longer time scales, gypsy moth populations go through host-pathogen population cycles, which in the gypsy moth and other forest defoliators lead to outbreaks that are terminated by the baculovirus (Moreau and Lucarotti 2007). The resulting predator-prey-type oscillations drive gypsy moth outbreaks at intervals of 5-9 years ((Dwyer et al. 2004), we neglect the effects of the gypsy moth fungal pathogen *Entomophaga maimaiga*, which was having only modest effects in our study areas in Michigan and Wisconsin, USA, when we collected our samples). Between insect outbreaks, virus infection rates are very low (Elderd et al. 2008), which may strengthen the effects of genetic drift.

In constructing models that allow for long-term dynamics, it is important to note that gypsy moths, like most outbreaking insects (Hunter 1991), have only one generation per year, and

therefore have only one epizootic per year. We thus nest our within-host/SEIR-type model into a model that describes host reproduction and virus survival after the epizootic (fig. 1). The SEIR model determines which hosts die during the epizootic and which virus strains killed those hosts. This information is used in a set of difference equations that describes the reproduction of the surviving hosts, the survival of the pathogen over the winter, and the evolution of host resistance, an important factor in gypsy moth outbreak cycles (Dwyer et al. 2000; Elderd et al. 2008).

By explicitly tracking the dynamics of individual hosts and pathogens, our model inherently includes the effects of genetic drift. If our models were to fail to explain the pattern of diversity seen in our data, other sources of evolution, such as mutation, migration, or pathogen natural selection, would have to be added to the models. Because we wanted to test whether genetic drift alone can provide a sufficient explanation of pathogen genetic data, however, we did not initially include these other evolutionary forces. Regardless, mutation rates are very low in double-stranded DNA viruses like baculoviruses (Rohrmann 2008), while previous work has shown that there are only minimal effects of spatial structure on the diversity of the gypsy moth baculovirus (Fujita 2007). In neglecting selection, we followed a standard approach from population genetics in which models that include only neutral forces are used as null hypotheses (Orr 1998), because sufficiently complex models of selection can explain almost any data.

To test whether the different sources of drift in our model are actually necessary to explain the data, we created two alternative models that simplified the effects of genetic drift, and we compared the predictions of all three models to data. For the first alternative model, we simplified the effects of genetic drift by assuming that the relative frequencies of different virus strains within a host did not change during pathogen population growth within hosts. To do this, we altered the model output such that the relative frequencies of virus strains released from a host upon host death were equal to the relative frequencies of virus strains just after

the transmission bottleneck, thereby eliminating the effects of replicative drift (Fig. 1). For the second alternative model, we further simplified the effects of drift by assuming that the relative frequencies of virus strains at the end of an infection were the same as their relative frequencies in the infectious cadaver that initiated the infection, thereby eliminating both replicative drift and transmission bottlenecks (fig. 1). Note that with the exception of virus strain frequencies, the host-pathogen dynamics are identical between these three models.

We tested our models using data on pathogen genetic variation across field-collected gypsy moth larvae. We collected larvae in outbreaking gypsy moth populations in Michigan and Wisconsin between 2000 and 2003, and we reared the larvae until they pupated or died of infection (Woods and Elkinton 1987). We used Illumina technology to sequence viral DNA from 143 infected larvae, and we mapped these reads to a reference genome (Kuzio et al. 1999) using "bowtie2" (Langmead and Salzberg 2012) (Supplemental Information). We summarized genetic variation using mean nucleotide diversity (Nei and Li 1979), the probability that two randomly selected alleles at a single nucleotide polymorphism ("SNP") are different (Supplemental Information).

## Results

The sequence data revealed substantial variation in the virus populations between hosts. We found 712 SNPs that segregated at the between host scale (alternative variant was the consensus in more than 4% of hosts), corresponding to approximately 0.4% of the genome. These 712 sites were polymorphic in some hosts but not others, which might occur if some hosts were exposed to multiple strains of virus, while others were exposed to only a single strain. The mean nucleotide diversity of the baculovirus averaged over these 712 sites ranged from 0.002 to 0.284 (mean$= 0.072$, s.d.$= 0.077$). Carrying out out the same calculation across the entire genome, while excluding these 712 sites, showed that polymorphism is vastly lower at the remaining

sites (range= $0.001$ to $0.003$, mean= $0.001$, s.d.$< 0.001$, Supplemental Information).

These patterns suggest that pathogen diversity within hosts is most likely acquired from the exposure of host insects to multiple virus strains rather than than from the generation of *de novo* mutations. If *de novo* mutation were creating high levels of diversity within hosts, we would expect to also see polymorphism spread across sites that were not polymorphic at the population level. Immune-system mediated diversifying selection is also an unlikely explanation, because insects lack clonal immune cell expansion (Vilmos and Kurucz 1998), and in any case, immune cell expansion would not explain why some hosts have substantially more pathogen diversity than others. Negative correlations between host families in susceptibility to different pathogen genotypes constitute yet a third unlikely explanation, because in the gypsy moth such correlations are almost always positive (Hudson et al. 2016). Migration of virus or infected larvae from nearby locations with different virus strains similarly cannot explain the data, because population structure in the gypsy moth virus is minimal (Fujita 2007).

Genetic drift, however, can explain the data, but only if we allow for effects of population processes at multiple scales of biological organization. To see why, we first describe the qualitative behavior of our model at each scale. We then show that only the model that includes effects of drift at multiple scales can explain the data.

Simulations of our within-host model show that the combination of transmission bottlenecks and replicative drift can substantially reduce pathogen diversity within hosts (fig. 2A-C). Demographic stochasticity, which can be seen as jaggedness in the lines, is strongest shortly after exposure when the pathogen population size is small. This stochasticity generates variability in the time to host death, and it also drives replicative drift. Similar effects occur in the corresponding linear birth-death model (Supplemental Information), but comparison of the two models makes clear that the immune system substantially slows the growth of the virus population early in the infection. By slowing pathogen population growth, the immune system

strengthens the effects of replicative drift.

Overwintered virus infects hatchlings during the initial emergence of hosts from eggs, an effect that is apparent in our simulations of the epizootic model (fig. 2D-E). After the overwintered virus decays, there is a short period when cadavers are rare, such that the vast majority of virus is present only within exposed larvae. When these larvae die, the virus that they release is transmitted to new larvae feeding on foliage. Note that the relative frequencies of different virus strains consumed by larvae during this time can fluctuate strongly due to the drift that occurs when cadavers are rare. Low densities of cadavers can thus alter the relative frequency of strains within hosts. In the figure, the initial host population consists of more than 10,000 hosts, reflecting the high densities at which baculovirus epizootics occur in insect populations in nature (Moreau and Lucarotti 2007). Demographic stochasticity nevertheless influences the composition of virus strains near the end of the epizootic, when the pathogen population begins to die out, in turn allowing drift to influence which virus strains cause infections within hosts.

Over longer time periods, fluctuations at the population scale (fig. 2F) produce host-pathogen cycles that match the dynamics of gypsy moth outbreaks in nature (Dwyer et al. 2000; Elderd et al. 2008). These large fluctuations can drive changes in the relative frequency of pathogen strains, especially when pathogen population sizes and overall infection rates are at their lowest in the troughs between host population peaks. Host-pathogen population cycles in our model thus further strengthen the effects of genetic drift on the pathogen.

Our combined model therefore shows that population dynamics across multiple scales provide opportunities for drift to act both within and between hosts, and at time scales ranging from hours to decades. To test whether population bottlenecks and replicative drift in particular have important consequences on pathogen genetic diversity, we compare the fit of the full model and the models lacking these factors to our sequence data.

This comparison shows that only the model that includes both replicative drift and bottlenecks can explain the data (fig. 3). The model that includes only drift at the host-population scale predicts within-host diversity levels that are much higher and much less variable than in the data. The model that includes population-scale drift and bottlenecks but not replicative drift correctly predicts that there will be substantial variation across hosts, but also predicts diversity levels that are much higher than in the data. The model that includes replicative drift and bottlenecks in contrast accurately predicts the entire distribution of diversity levels seen in the data.

This visual impression is strongly confirmed by differences in likelihood scores across models, averaged across realizations: model with neither bottlenecks nor replicative drift, log mean likelihood $= -503.2$; model with bottlenecks but not replicative drift, log mean likelihood $= -272.3$; model with both bottlenecks and replicative drift, log mean likelihood $= -64.7$. Because no parameters were fit to the diversity data, we do not need a model complexity penalty, but the difference in the number of parameters across models was in any case dwarfed by the differences in the likelihood scores. More broadly, because the model parameters were estimated entirely from experimental data on baculovirus infection rates (Supplemental Information), we are effectively carrying out cross-validation of the model.

The highly skewed distribution of nucleotide diversity apparent in our data can thus be explained by a model that allows for drift at multiple scales, and that includes multiple sources of drift within hosts, but not by simpler models. In addition, fig. 4 shows that the best model can reproduce entire distributions of diversity within individual hosts. Allele-frequency distributions in the model tend to have slightly shorter tails and narrower peaks compared to the data. These mild discrepancies may be partially explained by mutations that occurred during viral passaging or during library preparation, but they can also be explained by small biases introduced during the mapping of our short sequence reads to the reference genome

(Supplemental Information). The data therefore do not reject the model.

Our data were collected at times of peak or near-peak gypsy moth densities, which are the only times at which large numbers of larvae can be collected easily, and so the data do not directly show how changes in pathogen population size at the host-population scale affect pathogen variation. We therefore used our best model to explore how pathogen variation within hosts will change over the course of the gypsy moth outbreak cycle. Within-host diversity is predicted to be highest just as the host population begins to crash due to the pathogen, after which diversity is predicted to gradually decline until the next outbreak (fig. 5). Reductions in within-host variation due to transmission bottlenecks and replicative drift are thus counter-balanced by increases in within-host variation at the time of host population peaks, due to the high frequency of multiple exposures when host populations are large. Long-term, population-scale processes can thus also strongly affect within-host variation.

## Discussion

Our results show that a model that accounts for the effects of population processes at multiple scales can explain differences in pathogen variation across hosts in the gypsy moth baculovirus. Models that simplify the effects of genetic drift by ignoring effects of transmission bottlenecks and replicative drift in contrast cannot explain the diversity of this pathogen.

A basic prediction of population genetics theory (Nagylaki 1992), and a fundamental assumption of phylodynamic modeling (Grenfell et al. 2004), is that the effects of genetic drift are determined by population processes. Explicit tests of this assumption for infectious diseases, however, are rare. By using a model that was developed and parameterized using non-genetic data, our work shows that patterns of diversity in an insect pathogen can be explained by a drift model that accounts for population processes at multiple scales, but not by models that simplify the effects of drift. Our work thus provides clear evidence of the importance of population

processes in driving genetic drift in pathogens.

Models that invoke diversifying selection or host-specific immune selection might also provide reasonable explanations for our data, but such models are necessarily more complex than our model (Orr 1998). For our data, a drift-only model provides a sufficient explanation, but for other data, models that include more complex mechanisms may be necessary. Given that polymorphism has been widely observed in insect baculoviruses (Chateigner et al. 2015; Hodgson et al. 2001), our results suggest that baculoviruses present important opportunities for understanding the relationship between host-pathogen ecology and pathogen diversity.

We have shown that both population bottlenecks and replicative drift have detectable impacts on pathogen diversity. Due to the difficulty of separating these effects, previous studies of genetic drift have assumed that bottlenecks are complete (Ypma et al. 2013; Pennings et al. 2014; Klinkenberg et al. 2017), or have ignored key features of biology such as the nonlinear effects of host immune systems (Leonard et al. 2017), or have combined these sources of drift into a single parameter called effective population size $N_e$ (Volz et al. 2017). Distinguishing between transmission bottlenecks and replicative drift, however, may provide novel insights into disease control strategies. For example, the emergence of resistance to antibiotic drugs might be slowed if drug therapy windows are restricted to periods when the effects of replicative drift are strongest.

Whole-genome sequencing technologies have revolutionized our ability to measure pathogen variation, and have therefore been used in the early detection of drug-resistance (Mideo et al. 2016), in the discovery of novel viruses in nature (Lipkin and Anthony 2015), and in the diagnosis of disease in clinical settings (Wilson et al. 2014). Our work shows that whole-genome sequencing can also provide important insights into the ecology and evolution of host-pathogen interactions, especially when combined with nested disease models. The increasing availability of both parameterized models (Keeling and Rohani 2008) and genomic

data (Hatherell et al. 2016) suggests that our approach of using genetic data to challenge models of nested disease dynamics may be widely applicable.

## Data and Code Availability

All sequence data generated in this study are available through the Sequence Read Archive (SRA) of the National Center for Biotechnology Information (NCBI) under BioProject ID PRJNA386565.

All author-generated code used in this study is available from our GitHub repository: `https://github.com/dkenned1/KennedyDwyer`.

## Acknowledgments

Our work was supported by NIH grant R01-GM096655 to G Dwyer, V Dukic, and B Rehill. Computational support was provided by the Computing Research Institute and the Research Computing Center at the University of Chicago.

## References and Notes


ABEL, S., ZUR WIESCH, P. A., DAVIS, B. M., AND WALDOR, M. K. 2015. Analysis of bottlenecks in experimental models of infection. *PLoS Pathog* 11:e1004823.

ALIZON, S., LUCIANI, F., AND REGOES, R. R. 2011. Epidemiological and clinical consequences of within-host evolution. *Trends in Microbiology* 19:24–32.

ALIZON, S. AND VAN BAALEN, M. 2008. Acute or chronic? Within-host models with immune dynamics, infection outcome and parasite evolution. *The American Naturalist* 172:E244–E256.

ASHIDA AND BREY, P. T. 1998. Molecular Mechanisms of Immune Responses in Insects. Chapman & Hall, London.

AZARIAN, T., PRESTI, A. L., GIOVANETTI, M., CELLA, E., RIFE, B., LAI, A., ZEHENDER, G., CICCOZZI, M., AND SALEMI, M. 2015. Impact of spatial dispersion, evolution, and selection on Ebola Zaire Virus epidemic waves. *Scientific Reports* 5.

CHATEIGNER, A., BÉZIER, A., LABROUSSE, C., JIOLLE, D., BARBE, V., AND HERNIOU, E. A. 2015. Ultra deep sequencing of a baculovirus population reveals widespread genomic variations. *Viruses* 7:3625–3646.

DIDELOT, X., GARDY, J., AND COLIJN, C. 2014. Bayesian inference of infectious disease transmission from whole-genome sequence data. *Molecular biology and evolution* 31:1869–1879.

DWYER, G., DUSHOFF, J., ELKINTON, J. S., AND LEVIN, S. A. 2000. Pathogen-driven outbreaks in forest defoliators revisited: Building models from experimental data. *The American Naturalist* 156:105–120.

DWYER, G., DUSHOFF, J., AND YEE, S. H. 2004. The combined effects of pathogens and predators on insect outbreaks. *Nature* 430:341–345.

DWYER, G., ELKINTON, J. S., AND BUONACCORSI, J. P. 1997. Host heterogeneity in susceptibility and disease dynamics: Tests of a mathematical model. *American Naturalist* 150:685–707.

EAKIN, L., WANG, M., AND DWYER, G. 2014. The effects of the avoidance of infectious hosts on infection risk in an insect-pathogen interaction. *The American Naturalist* 185:100–112.

ELDERD, B. D. 2013. Developing models of disease transmission: insights from ecological studies of insects and their baculoviruses. *PLoS Pathog* 9:e1003372.

ELDERD, B. D., DUSHOFF, J., AND DWYER, G. 2008. Host-pathogen interactions, insect outbreaks, and natural selection for disease resistance. *The American Naturalist* 172:829842.

ELKINTON, J. S. AND LIEBHOLD, A. M. 1990. Population dynamics of gypsy moth in North America. *Annual Review of Entomology* 35:571–596.

FROST, S. D., PYBUS, O. G., GOG, J. R., VIBOUD, C., BONHOEFFER, S., AND BEDFORD, T. 2015. Eight challenges in phylodynamic inference. *Epidemics* 10:88–92.

FUJITA, P. A. 2007. Combining models with empirical data to examine dispersal mechanisms in the gypsy moth nucleopolyhedrosis host-pathogen system. Ph.d. dissertation, University of Chicago.

FULLER, E., ELDERD, B. D., AND DWYER, G. 2012. Pathogen persistence in the environment and insect-baculovirus interactions: disease-density thresholds, epidemic burnout and insect outbreaks. *The American Naturalist* 179.

GOG, J. R., PELLIS, L., WOOD, J. L., MCLEAN, A. R., ARINAMINPATHY, N., AND LLOYD-SMITH, J. O. 2015. Seven challenges in modeling pathogen dynamics within-host and across scales. *Epidemics* 10:45–48.

GRENFELL, B. T., PYBUS, O. G., GOG, J. R., WOOD, J. L. N., DALY, J. M., MUMFORD, J. A., AND HOLMES, E. C. 2004. Unifying the epidemiological and evolutionary dynamics of pathogens. *Science* 303:327–332.

GUTIÉRREZ, S., MICHALAKIS, Y., AND BLANC, S. 2012. Virus population bottlenecks during within-host progression and host-to-host transmission. *Current Opinion in Virology* 2:546–555.

HATHERELL, H.-A., COLIJN, C., STAGG, H. R., JACKSON, C., WINTER, J. R., AND ABUBAKAR, I. 2016. Interpreting whole genome sequencing for investigating tuberculosis transmission: a systematic review. *BMC Medicine* 14:1.

HODGSON, D. J., VANBERGEN, A. J., WATT, A. D., HAILS, R. S., AND CORY, J. S. 2001. Phenotypic variation between naturally co-existing genotypes of a Lepidopteran baculovirus. *Evolutionary Ecology Research* 3:687–701.

HUDSON, A. I., FLEMING-DAVIES, A. E., PÁEZ, D. J., AND DWYER, G. 2016. Genotype-by-genotype interactions between an insect and its pathogen. *Journal of Evolutionary Biology* 29:2480–2490.

HUNTER, A. 1991. Traits that distinguish outbreaking and nonoutbreaking macrolepidoptera feeding on northern hardwood trees. *Oikos* 60:275–282.

KEELING, M. J. AND ROHANI, P. 2008. Modeling Infectious Diseases. Princeton University Press, New Jersey.

KENNEDY, D. A., DUKIC, V., AND DWYER, G. 2014. Pathogen growth in insect hosts: inferring the importance of different mechanisms using stochastic models and response-time data. *The American Naturalist* 184:407–423.

KLINKENBERG, D., BACKER, J. A., DIDELOT, X., COLIJN, C., AND WALLINGA, J. 2017. Simultaneous inference of phylogenetic and transmission trees in infectious disease outbreaks. *PLoS Computational Biology* 13:e1005495.

KOELLE, K., COBEY, S., GRENFELL, B., AND PASCUAL, M. 2006. Epochal evolution shapes the phylodynamics of interpandemic influenza a (h3n2) in humans. *Science* 314:1898–1903.

KOT, M. 2001. Elements of Mathematical Ecology. Cambridge University Press, Cambridge.

KOUYOS, R. D., ALTHAUS, C. L., AND BONHOEFFER, S. 2006. Stochastic or deterministic: what is the effective population size of HIV-1? *Trends in Microbiology* 14:507–511.

KUZIO, J., PEARSON, M. N., HARWOOD, S. H., FUNK, C. J., EVANS, J. T., SLAVICEK, J. M., AND ROHRMANN, G. F. 1999. Sequence and analysis of the genome of a baculovirus pathogenic for *Lymantria dispar*. *Virology* 253:17–34.

LANGMEAD, B. AND SALZBERG, S. L. 2012. Fast gapped-read alignment with bowtie 2. *Nature Methods* 9:357–359.

LEE, R. S., RADOMSKI, N., PROULX, J.-F., LEVADE, I., SHAPIRO, B. J., MCINTOSH, F., SOUALHINE, H., MENZIES, D., AND BEHR, M. A. 2015. Population genomics of mycobacterium tuberculosis in the inuit. *Proceedings of the National Academy of Sciences* 112:13609–13614.

LEONARD, A. S., WEISSMAN, D., GREENBAUM, B., GHEDIN, E., AND KOELLE, K. 2017. Transmission bottleneck size estimation from pathogen deep-sequencing data, with an application to human influenza A virus. *Journal of Virology* pp. JVI–00171.

LIPKIN, W. I. AND ANTHONY, S. J. 2015. Virus hunting. *Virology* 479:194–199.

LORENZO-REDONDO, R., FRYER, H. R., BEDFORD, T., KIM, E.-Y., ARCHER, J., POND, S. L. K., CHUNG, Y.-S., PENUGONDA, S., CHIPMAN, J. G., FLETCHER, C. V., SCHACKER, T. W., MALIM, M. H., RAMBAUT, A., HAASE, A. T., MCLEAN, A. R. ., AND WOLINSKY, S. M. 2016. Persistent HIV-1 replication maintains the tissue reservoir during therapy. *Nature* 530:51+.

MALDARELLI, F., KEARNEY, M., PALMER, S., STEPHENS, R., MICAN, J., POLIS, M. A., DAVEY, R. T., KOVACS, J., SHAO, W., ROCK-KRESS, D., ET AL. 2013. HIV populations are large and accumulate high genetic diversity in a nonlinear fashion. *Journal of Virology* 87:10313–10323.

MIDEO, N., ALIZON, S., AND DAY, T. 2008. Linking within-and between-host dynamics in the evolutionary epidemiology of infectious diseases. *Trends in Ecology & Evolution* 23:511–517.

MIDEO, N., BAILEY, J. A., HATHAWAY, N. J., NGASALA, B., SAUNDERS, D. L., LON, C., KHARABORA, O., JAMNIK, A., BALASUBRAMANIAN, S., BJÖRKMAN, A., ET AL. 2016. A deep sequencing tool for partitioning clearance rates following antimalarial treatment in polyclonal infections. *Evolution, Medicine, and Public Health* 2016:21–36.

MOREAU, G. AND LUCAROTTI, C. J. 2007. A brief review of the past use of baculoviruses for the management of eruptive forest defoliators and recent developments on a sawfly virus in canada. *The Forestry Chronicle* 83:105–112.

NAGYLAKI, T. 1992. Introduction to Theoretical Population Genetics. Springer-Verlag, Berlin Heidelberg.

NEI, M. AND LI, W. H. 1979. Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Sciences of the United States of America* 76:5269–5273.

ORR, H. A. 1998. Testing natural selection vs. genetic drift in phenotypic evolution using quantitative trait locus data. *Genetics* 149:2099–2104.

PENNINGS, P. S., KRYAZHIMSKIY, S., AND WAKELEY, J. 2014. Loss and recovery of genetic diversity in adapting populations of HIV. *PLoS Genetics* 10:e1004000.

READ, A. F. AND TAYLOR, L. H. 2001. The ecology of genetically diverse infections. *Science* 292:1099–1102.

RENSHAW, E. 1991. Modeling Biological Populations in Space and Time. Cambridge University Press, Cambridge.

ROHRMANN, G. F. 2008. Baculovirus Molecular Biology. National Library of Medicine (US), Bethesda.

SAATY, T. L. 1961. Some stochastic-processes with absorbing barriers. *Journal of the Royal Statistical Society Series B-Statistical Methodology* 23:319–334.

TRUDEAU, D., WASHBURN, J. O., AND VOLKMAN, L. E. 2001. Central role of hemocytes in *Autographa californica M* nucleopolyhedrovirus pathogenesis in *Heliothis virescens* and *Helicoverpa zea*. *Journal of Virology* 75:996–1003.

VIGNUZZI, M., STONE, J. K., ARNOLD, J. J., CAMERON, C. E., AND ANDINO, R. 2006. Quasispecies diversity determines pathogenesis through cooperative interactions in a viral population. *Nature* 439:344–348.

VILMOS, P. AND KURUCZ, E. 1998. Insect immunity: Evolutionary roots of the mammalian innate immune system. *Immunology Letters* 62:59–66.

VOLZ, E. M., ROMERO-SEVERSON, E., AND LEITNER, T. 2017. Phylodynamic inference across epidemic scales. *Molecular Biology and Evolution* 34:1276–1288.

WEARING, H. J., ROHANI, P., AND KEELING, M. J. 2005. Appropriate models for the management of infectious diseases. *PLoS Medicine* 2:e174.

WILSON, M. R., NACCACHE, S. N., SAMAYOA, E., BIAGTAN, M., BASHIR, H., YU, G., SALAMAT, S. M., SOMASEKAR, S., FEDERMAN, S., MILLER, S., ET AL. 2014. Actionable diagnosis of neuroleptospirosis by next-generation sequencing. *New England Journal of Medicine* 370:2408–2417.

WOODS, S. A. AND ELKINTON, J. S. 1987. Bimodal patterns of mortality from nuclear polyhedrosis-virus in gypsy-moth (*Lymantria-dispar*) populations. *Journal of Invertebrate Pathology* 50:151–157.

YPMA, R. J., VAN BALLEGOOIJEN, W. M., AND WALLINGA, J. 2013. Relating phylogenetic trees to transmission trees of infectious disease outbreaks. *Genetics* 195:1055–1062.

ZWART, M. P., HEMERIK, L., CORY, J. S., DE VISSER, J. A. G. M., BIANCHI, F. J. J. A., VAN OERS, M. M., VLAK, J. M., HOEKSTRA, R. F., AND VAN DER WERF, W. 2009. An experimental test of the independent action hypothesis in virus-insect pathosystems. *Proceedings of the Royal Society of London Series B-Biological Sciences* 276:2233–2242.

**Figure 1:** Schematic of the nested model. Reading from bottom to top, the host population size $N_g$ and the infectious cadaver population size $Z_g$ in generation $g$ depend on host and pathogen population sizes in generation $g-1$ and the disease dynamics in that generation. Moving up, the disease dynamics in generation $g-1$ follow a stochastic "SEIR" model (Keeling and Rohani 2008), such that a susceptible host $S_i$ becomes exposed to cadaver $P_k$ at rate $\nu_i q$, where $\nu_i$ is the risk of exposure for host $i$ and $q$ is the probability of death given exposure, which arises from the within host virus dynamics. Similarly $p(\tau)$, which is the probability of a host dying from virus infection at time $\tau$ post exposure, is determined by the dynamics of the pathogen within a host. $q$ is related to $p(\tau)$ in that $q = \int_0^\infty p(\tau)d\tau$. Within a host, virus particles $x$ can reproduce, or interact with immune cells $y$, resulting in the removal of both the virus particle and the immune cell. An infection fails to kill the host if all virus particles are cleared $x = 0$, but it dies if the total number of virus particles reaches an upper threshold $C$. Further details are in the Supplemental Information. Models that lack replicative drift assume that the frequency of a particular virus strain at time of death $f_l(\tau)$ is equal to the frequency of that strain immediately after the time of exposure $f_l(0)$. The model that lacks transmission bottlenecks assumes that the number of copies of a particular virus strain at the beginning of an infection $x_l(0)$ is equal to the total number of virus particles that invade the host $\sum_l x_l(0)$ times the relative frequency of that virus strain in the cadaver that caused exposure $\frac{P_l}{\sum_p P_p}$.

**Figure 2:** Simulations of the nested model. In all panels (A)-(F), colored curves represent the pathogen population size for different virus strains, and the black curve shows the total pathogen population size. The colored bar at the top of each panel shows the relative frequencies of each virus strain over time. Panels (A)-(C) show three realizations of the within-host virus growth model. The top colored panel left of time $0$ shows the frequency of virus strains in a cadaver that a host was exposed to at time $0$. In the model, death occurs when

the total number of virus particles within a host hits an upper threshold. To aid visualization, here we set the pathogen population size at host death to be $10^4$, as opposed to the more realistic value of $10^9$ that we use when comparing our models to data. The time of death differs between simulations due to demographic stochasticity in virus growth, and in each simulation it is marked by a dotted, vertical line. Panels (D) and (E) show two realizations of our stochastic SEIR-type epizootic model starting from identical initial conditions. Note that the curves here show cadaver quantities, rather than virus particles as in panels (A)-(C). Epizootics are seeded by overwintered cadavers that infect emerging larvae. As these cadavers decay, total cadaver quantity drops to low levels, such that the pathogen population is almost entirely composed of virus particles inside living hosts. These hosts then die seeding future rounds of infections. Panel (F) shows a realization of our between-generation pathogen model. The curves here denote the total number of virus-killed hosts in each generation. Note that the frequency of pathogen strains can drift over time, an effect that is particularly noticeable during troughs of infection.

**Figure 3:** Comparison of the predictions of our models (gray-shaded areas, showing 95 percent confidence intervals of model realizations) to the distribution of nucleotide diversity within 143 individual infected hosts calculated from our sequence data (black dots, points are based on nucleotide diversity within hosts). (A) shows the predictions of a model that lacks both transmission bottlenecks and replicative drift, (B) shows the predictions of a model that includes transmission bottlenecks but not replicative drift, and (C) shows the predictions of a model that includes both transmission bottlenecks and replicative drift.

**Figure 4:** Representative distributions of allele frequencies from individual hosts in our best model (A-E) and in our data (F-J). Each plot shows the distribution of allele frequencies within

a single individual at 712 SNPs, showing only the frequency of the most common allelic variant at each locus within that host. The number on each plot is the mean nucleotide diversity within that particular host. Model plots are aligned with similar data plots.

**Figure 5:** Model predictions of the effects of changes in the populations of susceptible and infected hosts on within-host pathogen diversity, over the host-pathogen population cycle. (A) Changes in the population size of uninfected hosts at the start of a generation. (B) Changes in the population of infectious cadavers at the start of a generation (blue) and in mean nucleotide diversity of newly created cadavers in that generation (red).

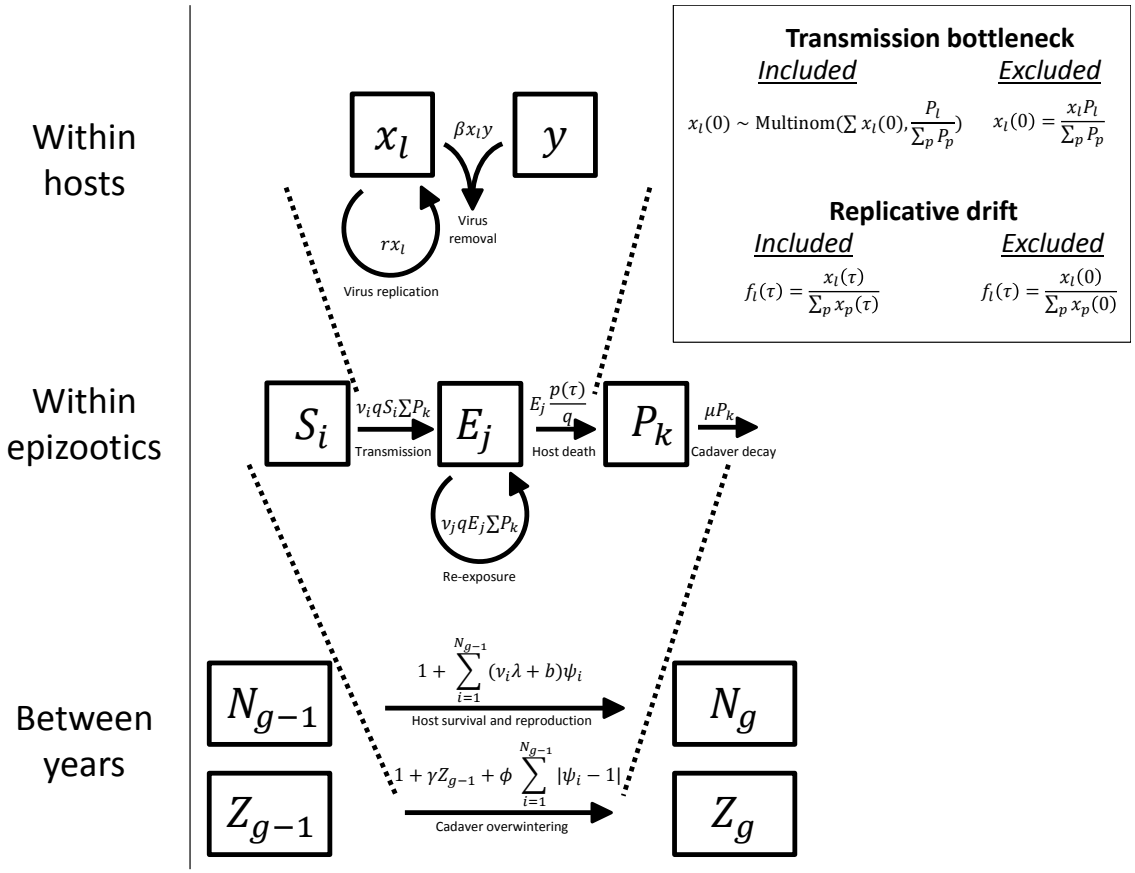
Within hosts
Within epizootics
Between years
$x_l$
$\beta x_l y$
$y$
$r x_l$
Virus removal
Virus replication
$S_i$
$\nu_i q S_i \sum P_k$
Transmission
$E_j$
$E_j \frac{p(\tau)}{q}$
Host death
$P_k$
$\mu P_k$
Cadaver decay
$\nu_j q E_j \sum P_k$
Re-exposure
$N_{g-1}$
$1 + \sum_{i=1}^{N_{g-1}} (\nu_i \lambda + b)\psi_i$
Host survival and reproduction
$N_g$
$Z_{g-1}$
$1 + \gamma Z_{g-1} + \phi \sum_{i=1}^{N_{g-1}} |\psi_i - 1|$
Cadaver overwintering
$Z_g$
Transmission bottleneck
Included
Excluded
$x_l(0) \sim \text{Multinom}(\sum x_l(0), \frac{P_l}{\sum_p P_p})$
$x_l(0) = \frac{x_l P_l}{\sum_p P_p}$
Replicative drift
Included
Excluded
$f_l(\tau) = \frac{x_l(\tau)}{\sum_p x_p(\tau)}$
$f_l(\tau) = \frac{x_l(0)}{\sum_p x_p(0)}$

**Figure 1**

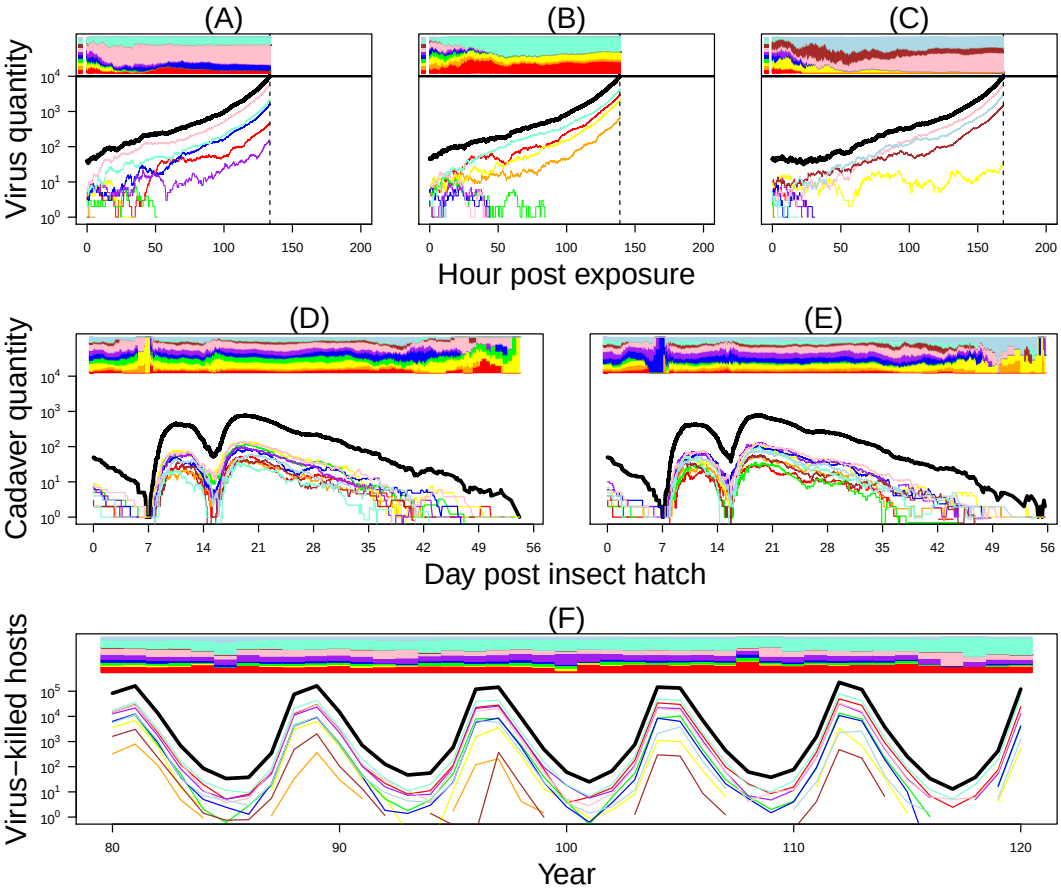
(A)
(B)
(C)
Virus quantity
Hour post exposure
(D)
(E)
Cadaver quantity
Day post insect hatch
(F)
Virus−killed hosts
Year

**Figure 2**

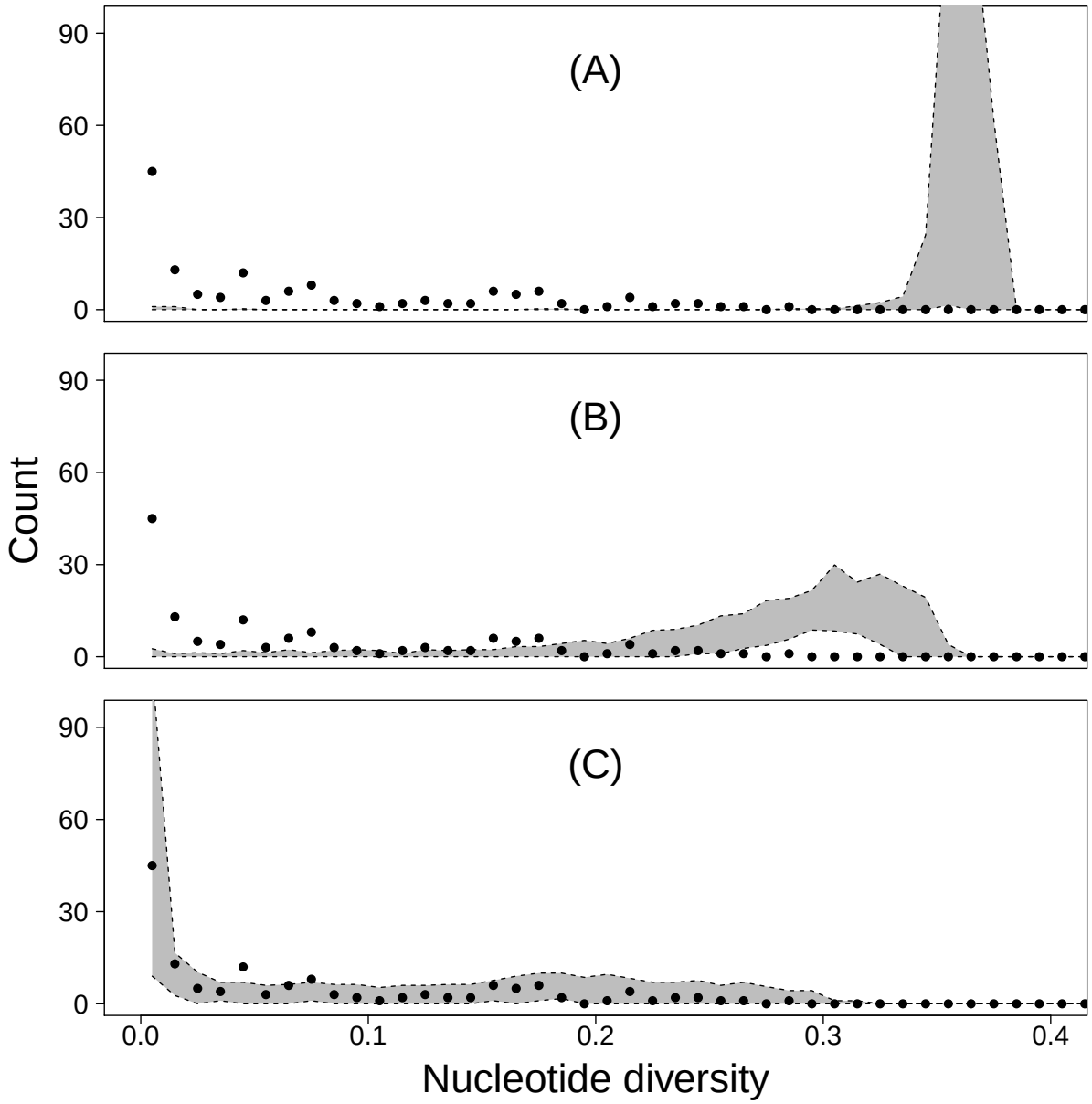


**Figure 3**

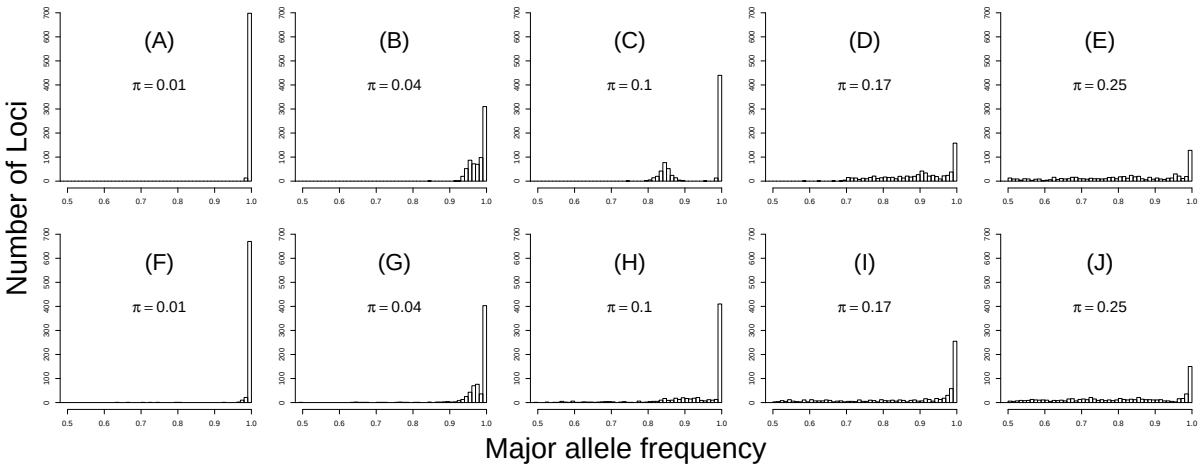
(A)
π = 0.01
(B)
π = 0.04
(C)
π = 0.1
(D)
π = 0.17
(E)
π = 0.25
(F)
π = 0.01
(G)
π = 0.04
(H)
π = 0.1
(I)
π = 0.17
(J)
π = 0.25
Number of Loci
Major allele frequency

**Figure 4**

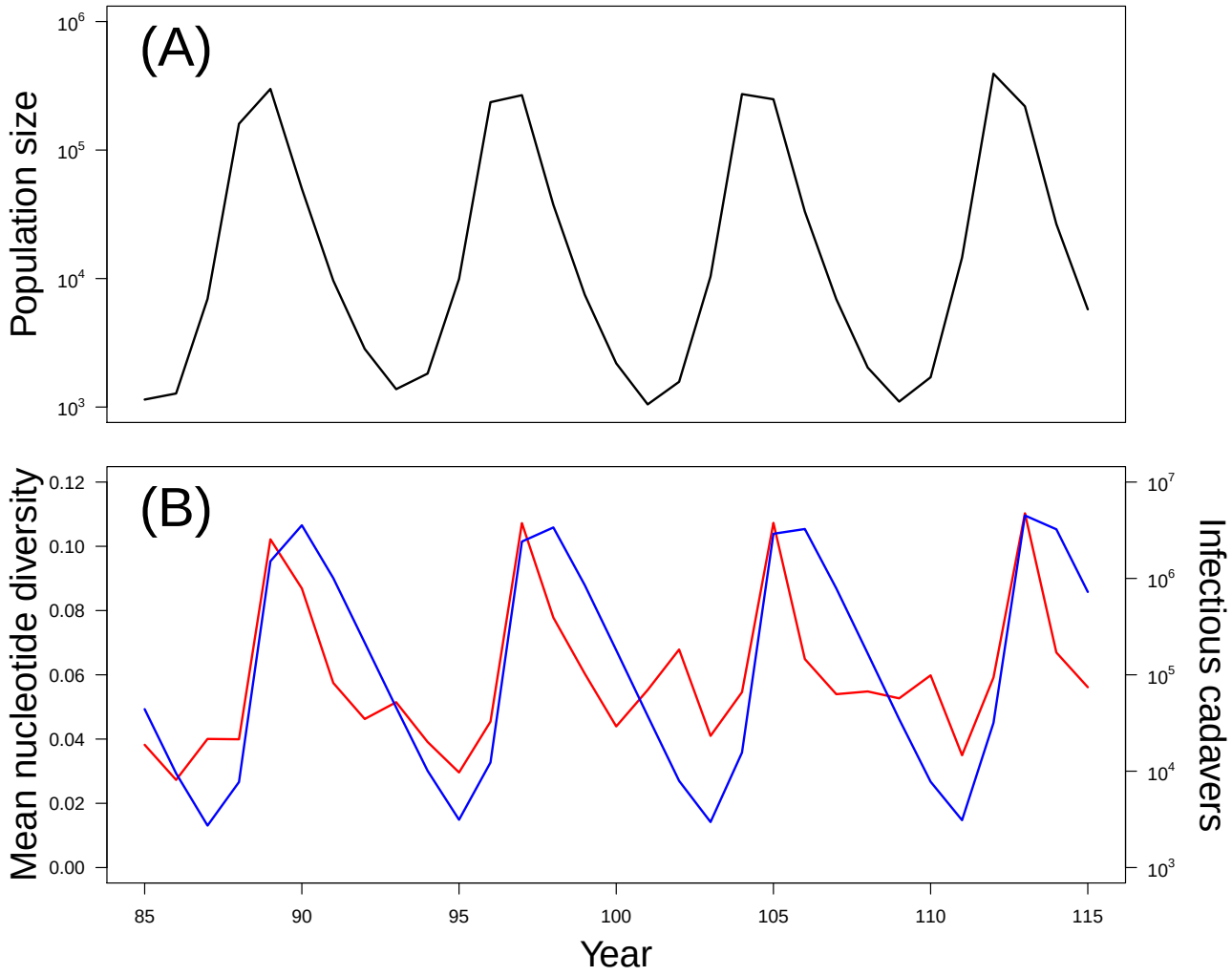
(A)
Population size
$10^6$
$10^5$
$10^4$
$10^3$
(B)
Mean nucleotide diversity
0.12
0.10
0.08
0.06
0.04
0.02
0.00
Infectious cadavers
$10^7$
$10^6$
$10^5$
$10^4$
$10^3$
85
90
95
100
105
110
115
Year

**Figure 5**