Modeling the evolution of segment trees reveals deficiencies in current inferential methods for genomic reassortment

Qianying Lin[1,†], Emma E. Goldberg[1], Thomas Leitner[1], Carmen Molina-París[1], Aaron A. King[2,3], and Ethan O. Romero-Severson[1,†]

[1]Theoretical Biology and Biophysics, Los Alamos National Laboratory, Los Alamos NM, USA

[2]Center for the Study of Complex Systems and Departments of Ecology & Evolutionary Biology and Mathematics, University of Michigan, Ann Arbor MI, USA

[3]Santa Fe Institute, 1399 Hyde Park Road, Santa Fe NM, USA

[†]Correspondence: *qianying@lanl.gov*, *eoromero@lanl.gov*

September 20, 2023

Abstract

Reassortment is an evolutionary process common in viruses with segmented genomes. These viruses can swap whole genomic segments during cellular co-infection, giving rise to new viral variants. Large-scale genome rearrangements, such as reassortment, have the potential to quickly generate new phenotypes, making the understanding of viral reassortment important to both evolutionary biology and public health research. In this paper, we argue that reassortment cannot be reliably inferred from incongruities between segment phylogenies using the established remove-and-rejoin or coalescent approaches. We instead show that reassortment must be considered in the context of a broader population process that includes the dynamics of the infected hosts. Using illustrative examples and simulation we identify four types of evolutionary events that are difficult or impossible to reconstruct with incongruence-based methods. Further, we show that these specific situations are very common and will likely occur even in small samples. Finally, we argue that existing methods can be augmented or modified to account for all the problematic situations that we identify in this paper. Robust assessment of the role of reassortment in viral evolution is difficult, and we hope to provide conceptual clarity on some important methodological issues that can arise in the development of the next generation of tools for studying reassortment.

***Key words*:** phylodynamics, genomic reassortment, molecular epidemiology, population genetics

Introduction

Reassortment is the evolutionary process by which viruses with segmented genomes exchange genetic material [1, 2]. When a cell becomes dually infected, that is, infected by two distinct viruses, the resulting viral offspring can contain a mixture of genomic segments from each parent [3]. The genetic diversity and fitness of segmented viruses such as influenza A virus [4, 5], rotaviruses [6, 7], and bunyaviruses [8, 9] are influenced by reassortment. For example, in bunyaviruses, the M segment encodes the viral glycoproteins that mediate both cell fusion and immune responses [10]. By swapping M segments, therefore, viruses may gain access to new hosts. As a concrete example, the hemorrhagic Ngari virus emerged from reassortment between the only mildly pathogenic Bunyamwera and Batai viruses [11, 12]. Understanding the molecular mechanism, the evolutionary process, and the epidemiological implications of reassortment is essential for anticipating and combating emerging infectious diseases [13], improving public health strategies [14], and guiding the development of novel diagnostics, therapeutics, and vaccines [5].

Reassortment can be thought of as a special case of recombination, one in which the breakpoints are fixed and correspond to the endpoints of each segment [3]. In population genetics, a phylogenetic tree is commonly used to represent evolutionary history over time, in which the root represents the most recent common ancestor (MRCA) of all the sampled genomes, and internal nodes represent a splitting of one lineage into two, denoting the ancestral relationships between sampled tips [15]. For viral pathogens, this pattern of branching is reflective both of the epidemiology of who infected whom, and also exactly who within the population was sampled [16]. The logic underlying reconstructions of reassortment is that all else being equal, topological discordance between segment trees is caused by historical reassortment events (Fig. 1). For example, the observation that two viruses that co-cluster in a tree inferred from one genomic segment but are quite distant in a tree inferred from another segment is likely the result of a recent reassortment event. This logic lays the foundation of many reassortment inferential methods, which are implemented in one of two different ways. One approach begins with segment trees and then attempts to synthesize their topological incongruence into an Ancestral Reassortment Graph (ARG), wherein specific reassortment events correspond to reticulations, i.e., a lineage originated from the partial merging of two ancestor lineages [17–20]. This is analogous to the practice of reconciling gene trees to infer recombination [21, 22]. These methods are intuitively appealing but they cannot accurately pinpoint the timing of reassortment events in dated segment trees [23]. A second approach seeks to infer an ARG—or in the case of Ref. 24, a very closely related data structure called a phylogenetic network—jointly with the segment trees using the likelihood under a coalescent-with-reassortment model [24, 25]. These approaches generally oper-

ate as an extension of the Kingman coalescent [26, 27], by only considering the events in the population that are specifically ancestral to the sampled genomes. In this coalescent-based formulation, the likelihood of an ARG is computed backward in time (from tips to the root), wherea coalescent event collapses two existing lineages into one and a reassortment event 'creates' a new lineage from an existing one. This approach is analogous to the multi-species coalescent models used for joint gene tree–species tree inference [28, 29]. This second approach is the current gold standard in that it allows for the identification of specific reassortment events in time and has an established statistical basis [24]. However, both approaches share the core assumption that the best path to inference of reassortment events is via an ARG-like object, itself derived from incongruities in the segment-trees. We refer to these as 'incongruence-based' methods.

Here, we argue that all incongruence-based methods are fundamentally limited by the fact that any set of segment trees, or their corresponding ARG, is consistent with multiple and significantly different patterns of reassortment (Fig. 2). Consequently, different methods are needed. Instead of an incongruence-based approach that works backward in time along the observed segment trees, we start by formulating a population process that explicitly includes reassortment based on the theory of Markov Genealogy Processes [30], working forward in time to document the full history of infection, reassortment, recovery, and sampling in the entire population. We use this formalism to document the types and frequencies of situations that can occur when relying on incongruence-based approaches, and we define the missing information that is required to correctly infer the presence and timing of genomic reassortment.

Methods

Model overview

Our model defines a population process that consists of both the epidemiology of the host population and the ancestry of the viral population. Individuals can become infected and, once infected, can become infected again. We make the simplifying assumption that subsequent infection is resolved instantly, such that an individual only hosts one viral genotype at a time. We reserve the term *re-infection* for the situation in which the new viral genotype within the host differs from the previous one. To keep the model simple, we define reassortment to mean the combination of re-infection, cellular co-location, reassortment of one or more genomic segments, and dominance of the new reassortant at the host level. Each subset of genomic segments can, in principle, have a unique reassortment rate due to differences in within- or between-host survival, or the biophysical aspects of viral particle formation. Infected individuals leave the system by either dying or becoming sampled. For simplicity, we assume that sampling is destructive to avoid the important but out-of-scope issues that arise in sampling direct descendants of already-sampled tips in phylogenetic trees. All

of these events are governed by the parameters of a stochastic process. We assume that all individuals are exchangeable, so there is no host- or virus-level heterogeneity such as different behaviors or fitness. Further details on the model construction are provided below and in the Supplementary Materials.

We use the term *population history* to refer to the full history of all events that update the host dynamics or the genealogical structures in a population. In the context of our model, this is a record of who infected whom when, and the times of all other sampling, removal, and reassortment events. Figure 1 illustrates several important subsets of the population history. The *infection history* is a time-scaled directed graph, showing which individual infected (or re-infected) which other individual, and at what time. The *first-infection tree* is the subset of the infection history that only includes the first-infection event (i.e., the creation of a newly infected individual) for each infected and sampled individual, without the specific information of who infected whom at the internal nodes. An infection history defines exactly one first-infection tree, but one first-infection tree is consistent with many possible infection histories. The *segment trees* are the observed genealogies of the individual viral segments in a host population. We assume that the true genealogy is directly observable, though in reality, this involves inferring a time-scaled phylogenetic tree. We make no attempt to deal with the intricacies of tree reconstruction since the difficulties with incongruence-based reassortment inference can be made evident without doing so.

Modeling segment genealogies with reassortment using the Markov Genealogy Process

We implemented our simulation model as a Markov Genealogy Process (MGP) [30] containing two distinct levels of biological processes: epidemiology of the host population and ancestry within the viral population (Table S1). A conceptual description is provided here, and a more complete technical description is available in the Supplementary Materials and Ref. [30]. The minimal model we use in this paper has six per-lineage rates, since we restrict ourselves to the case where the segmented virus has two different segments, denoted A and B: the first-infection rate, λ; the recovery rate, μ; the sampling rate, ψ; the reassortment rates for segment A and B, ρ_A and ρ_B, respectively; and the joint reassortment rate, ρ_{AB}, accounting for the displaced segment trees when compared with the first-infection tree, as shown in option (ii) of Fig. 2(e). Note that the effective removal rate of an individual is $\mu + \psi$, as we assume infected individuals who got sampled are immediately hospitalized or quarantined, and thus removed from the infected population.

The spread of infection in the host population is described by stochastic exponential growth at rate $\lambda - (\mu + \psi)$. The rates of re-infection are governed by segment-specific, compound parameters that give the rate of re-infection leading to a successful reassortment event, which we refer to overall as the reassortment rates. The MGP records the full infection history (identity of source and recipient, and time of infection)

in a format that enables much of the downstream computations that we use to count types of reassortment inference situations. The first-infection tree is a subset of infection history that includes only the sub-tree of sampled tips and the nodes corresponding to the time they were first infected; source and recipient identity are also removed.

Within the infection history, we track the genealogy of each of the segments of the viral genome (Fig. 1c). For simplicity, we ignore viral diversification within each host, so nodes in the segment genealogies corresponding to infection events occur exactly at the time of infection. In the absence of reassortment, all segment trees will be identical to the first-infection tree. In the case of a reassortment event, we assume that the within-host diversity immediately collapses to the relevant reassorted genomic segment (Fig. 1d). In terms of our model, this involves a coupled death-birth event where the recipient node is removed (along with its branch), and then reattached at the source's node with a zero-length branch in the relevant segment tree (Fig. S1). This creates a new node in the population history, but the population size remains unchanged.

The descriptions above apply to the entire population of hosts and the viruses that infect them, i.e., the full population process. Not all individuals will be observed, however, and the sampled genealogy only represents the ancestral relationships between viral genomic samples. In the MGP, the full infection history, first-infection tree, and segment trees are pruned to include only the samples taken. Note that the MGP need not assume that sampling is sparse or has any other non-pathological restrictions. Samples are taken sequentially from the infected population at the per-lineage sampling-and-removal rate, ψ. For visual simplicity, figures illustrating stylized population histories or trees are always shown as if a homochronous, batch sample was taken at a given time.

Simulating infection histories and segment genealogies

The MGP provides a formal mathematical framework for the events described above. In this study, we use it as a means to run forward-in-time stochastic simulations that record the full population history of first-infections, recoveries, reassortments, and sampling events. We thus simultaneously simulate the host epidemiological process and the genealogies for all viral genomic segments. In this paper, we will only consider two reassorting segments, but the MGP has no restriction on the number of genomic segments or their corresponding reassortment rates.

For the host population dynamics, we use the linear-death-birth process [31, 32] at an asymptotically constant population size. We start from one single infected individual ($I = 1$), while setting the first-infection rate $\lambda = 200$ and shutting down recovery, sampling, and reassortment until the population size reaches $I = 100$. We then turn on reassortments (i.e., $\rho_A > 0$ or/and $\rho_B > 0$, $\rho_{AB} = 0$) and run a burn-in period where

first-infection rate equals to the recovery rate and sampling is kept shut down (i.e., $\lambda = \mu = 5$ and $\psi = 0$), before proceeding with sampling (i.e., $\lambda = \mu + \psi$, where $\mu > 0$ and $\psi > 0$). We run the simulation until 50 samples are obtained. This approach guarantees that the population size stays asymptotically constant at 100 during the studied period of the obtained trees.

Sequence data simulation and reassortment inference

We first simulated the full population history at an asymptotically constant population size ($I = 100$) with rates $\lambda = 5$, $\mu = 4.75$, $\psi = 0.25$, $\rho = \rho_A + \rho_B + \rho_{AB} = 0.2$, until 50 samples were obtained. For each situation discussed in the results below, we produced 100 different replicates. Each segment genealogy represents precisely the history of that segment, but inferring a segment tree in practice requires genetic sequence data. Therefore, we used IQ-TREE 2 [33] to simulate alignments with a sequence length of 10^3 nucleotides using the Jukes-Cantor model for each tree, at an evolutionary rate 5×10^{-3} per site [24].

We analyzed our simulated data with two leading software packages for inferring reassortment. To use the `CoalRe` [24] phylogenetic network model, we imported the alignments into BEAST2 [34] and used the same mutation model. `TreeKnit` [19] only requires the Newick format of tree data, so we directly input each pair of segment trees.

Results

Reassortment is a type of coupled death-birth event

Reassortment is a complex process that involves cellular co-infection, competition among different viral strains, host immune responses, and the biophysics of genome packaging and virion formation. Each host-virus system will be unique in ways that may affect inferences regarding, for example, the timing of reassortment events. However, taking out the complex biology and within-host dynamics, the most basic operationalization of reassortment in a model is that of a coupled death-birth process. This process is illustrated in Fig. S1 for a hypothetical virus with two segments, A and B. In that case, host 0 co-infects host 2, swapping the B segment. Host 2's original B segment is lost in the population and their original branch in segment B is dissolved (i.e., the original branch 'dies'). Correspondingly, host 2 now has a segment B that is directly descended from host 0 (i.e., a 'birth' occurs in segment B at host 0). This type of coupled death-birth event is how reassortment produces differences in both topology and branch lengths in segment trees. Understanding the abstract effect of reassortment on genomic segment trees as this type of coupled death-birth event, in the context of a broader population process, is central to understanding the situations we raise below.

The correct number of *visible* reassortment events cannot be computed from segment trees alone

We can define different relevant types of reassortment with respect to their influence on the data (illustrated in Fig. S2). We call a reassortment event *visible* if it had some influence on the observed data. Alternatively, a reassortment event is *invisible* if it is not ancestral to the sampled genomes, or if it is ancestral but was replaced by a reassortment event that happened later in the same lineage. In this section, we focus on counting the number of visible reassortment events for a set of segment trees. That is, 'what is the number of reassortment events needed to explain the data?'.

The common way of counting the number of visible reassortment events is to find the minimum number of 'remove-and-rejoin' operations [18, 23, 35] that are needed to reconcile the segment trees. The idea is to resolve any structural differences between segment trees by removing a tip or clade in one tree, along with its ancestral branch, and then rejoining that branch somewhere else in the same tree so as to reduce the differences between these two trees. Figure S4 illustrates this idea. Generally, for such an approach the governing principle is parsimony: to propose the minimum number of reassortments that are needed to transform one segment tree into the other. We denote the number of reassortments estimated by this method as the *observed minimum* number of reassortments (Fig. 8). Note that when we talk about 'structural differences' between segment trees, we are referring to not only the tree topologies but also the branch lengths, which makes sense under the assumption that the segment trees can be observed without error. In practice, counting the number of reassortments might only rely on the inferred phylogenetic topology, such that any branch length difference is due to the difference in mutation rate between segments. In that case, the remove-and-rejoin method is identical to the subtree-prune-regraft method [35]. However, we argue that, because reassortment can cause differences in branch lengths alone, considering only evolutionary rate heterogeneity or stochasticity as the source of branch length variation, biases both the estimation of evolutionary rate heterogeneity and reassortment.

The problem that arises in applying remove-and-rejoin on the segment trees is that, even when applied correctly, it does not necessarily find all the reassortment events that are visible in the data when one also considers the underlying population history. Figure 2 illustrates how the exact same set of segment trees can have a different number of visible reassortment events, depending on the infection history. While it is possible for a single remove-and-rejoin move (in this example) to convert one segment tree into the other, this resolution is not biologically possible under some infection histories. Existing incongruence-based methods that consider only segment trees will thus tend to under-report reassortment events.

We argue that finding the number of visible reassortment events in the observed data does not require knowledge of the full infection history; rather, the first-infection tree alone is sufficient. The first-infection tree can be thought of as the tree structure that would have been observed in the segment trees if there had been no reassortment. It therefore represents a starting point for calculating the number of reassortment events required to obtain each of the segment trees (Fig. S4). Without the first-infection tree, the estimated number of reassortment events will likely be lower than the true number of visible reassortments unless one of the segment trees has no visible reassortment. This exception could be assumed only if the true population process governing reassortment in some system never allows reassortments in one of the segments, and if it is known which segment that is (if there are more than two segments).

Four situations caused by reassortment that create problems for existing methods

Using the same population process described above, we now show how attempts to estimate other quantities of the reassortment history (e.g., reassortment rate and reassortment times), beyond the number of visible events, using incongruence-based approaches run into problems in specific situations. We identify four situations that we label 'invisible', 'inaccurate', 'reversed', and 'obfuscated'. We address each situation in turn below, identifying where the potential for error arises and the impact that it might have on reassortment inference. We believe any method claiming to infer reassortment events should be able to explicitly model, and account for, these situations for it to be considered robust.

Figure 3 illustrates the situation of *invisible* reassortment. This is the most straight-forward and easiest to correct situation. Invisible reassortment occurs when either a reassortment occurs in the part of the population that is unsampled, or when it is overwritten by another reassortment event that happens later (see Fig. S2). In the context of a structural analysis of reassortment, this is not viewed as problematic inasmuch as such analyses attempt to identify the minimum number of reassortments which explain the data and would not claim to detect invisible reassortments. However, for methods that attempt to infer the reassortment rate, failure to correct this situation will bias the estimation of the reassortment rate. In Fig. 3 the invisible reassortment shown is not ancestral to the sample and, therefore, will only directly influence inferences concerning the broader effect of reassortment in the population (e.g., the distribution of fitness effects caused by any reassortment). However, invisible reassortment also occurs along sampled lineages (i.e., ancestral to the sample). Along a given lineage and its descendants, we can only observe the most recent successful reassortment event(s). Figure S3 illustrates how we dismiss those invisible reassortments along the lineages before performing the remove-and-rejoin moves to resolve the structural differences between trees. If a method does not correctly integrate the probability that an inferred ancestral reassortment event is

visible, it could lead to an underestimation of the reassortment rate or artifactual support for heterogeneous reassortment rates. This is a particularly acute problem for methods based on the coalescent, as they condition on the observed data and are unable to resolve this problem as the sampling probability in such models is not defined.

Figure 4 illustrates the situation of *inaccurate* inference of reassortment times that occurs when both lineages (or one of their direct descendants) involved in a reassortment are sampled. At time t_B, there is a reassortment event between samples 1 and 4, and, because both individuals are eventually sampled, the node corresponding to that event shows up as the most recent common ancestor of individuals 1 and 4 in the segment B tree. These types of nodes pose a problem for methods based on the manipulation of an ARG in the context of a reverse-time coalescent process. The situation is that in the coalescent-with-reassortment model, the coalescent and reassortment process must be assumed to be independent to ensure that the coalescent prior makes sense. However, as is clear in this example and from the representation of reassortment events in our model, an internal node in one or more of the sampled segment trees can be caused by a reassortment event rather than a first-infection event. In Fig. 4, the node R^B corresponds to a reassortment event that produced a node in segment tree B, where both children of that node (individuals 1 and 4) were sampled leading to the node associated with the reassortment event being in the segment B tree. To explain this, a coalescent-based approach needs to insert two nodes into the ARG, one corresponding to the reassortment event to split and another separate node for that additional lineage to coalesce. Because both nodes cannot co-occur, there must be a bias introduced. In principle, the distance between these two nodes can be made arbitrarily small such that they effectively occur at the same time. However, this is unlikely in practice unless the data are highly informative (i.e., very long, clean sequences). In the more common case where the data cannot entirely overcome the priors, the distance between the two nodes will be influenced by the coalescent prior which, by definition, will have some non-zero mean, encouraging bias in the estimation of the reasortment and coalescent times.

Figure 5 illustrates the situation of *reversion* in the inference of reassortment. Reversion occurs when a segment is passed between sampled individuals, but only one offspring of the reassortment event is sampled, leaving the node corresponding to the reassortment event to be unobserved in the segment tree (i.e., individuals 5 and 6 are unsampled in Fig. 5C). The statistical situation of reversion has a similar effect to that in inaccuracy, in that either one or both of the timing of the reassortment and corresponding coalescent events in an ARG will not be correct. The cause is slightly different, however. When only one lineage descending from a reassortment event is sampled, the corresponding ARG requires a tri-furcation (three lineages coming together in one node) at the sampled parent node of an unsampled reassortment event. Generally speaking,

if some or all of the nodes of the ARG are assumed to follow a coalescent prior, then trifurcations will not be allowed, creating a similar situation as in the case of inaccuracy, where additional, artifactual nodes must be inserted into the ARG (nodes A' and C' in Fig. 5D).

Figure 6 illustrates the situation of *obfuscation* in the inference of reassortment. The defining aspect of obfuscation is that all segment trees have one or more reassortment events and at least one of these events is only partially sampled (i.e., only one of the descendants is sampled). We call this situation obfuscation because the first-infection tree no longer has the same structure as one of the segment trees. The effect of obfuscation is that the data appear to be explainable by fewer than the number of visible reassortments and the underlying population history is thus completely obscured.

The situations that cause problems for the inference of *visible* reassortment events are common

Of the four situations that we discussed above, three are potential problems for inferring visible reassortments. In this section, we assess the frequency at which those situations arise both theoretically and in practice. We do not assess the frequency of invisible reassortment events as almost all reassortment events will be invisible, even in the context of an extremely large sample, and invisible reassortment events are not a problem in the assessment of visible reassortment events per se.

Using our population model we simulated a full population history and extracted the sampled segment trees for a grid of different parameter combinations, then counted the frequency of the different types of situations (details in the Supplementary Materials). Figure 7 shows the frequency of each situation for a range of simulation parameter values. The outcome variable is the proportion of nodes that have the type of referred situation averaged over 100 simulations. The *inaccuracy* situation increases as the first-infection rate becomes small with respect to a fixed sampling rate. This makes sense as inaccuracy occurs when both descendants from a reassortment event are sampled, which is more likely when the first-infection rate is low, meaning that a larger proportion of nodes in the tree are caused by reassortment events, rather than first-infection events. Both *reversion* and *obfuscation* follow a similar pattern of being less frequent when the first-infection rate is high and the reassortment rate is low with respect to the sampling rate, and more frequent when the first-infection rate is low and the reassortment rate is high. Both of these types of situation occur when the nodes caused by reassortment events are not fully sampled (i.e., they do not appear as a node in the sampled segment trees), which will occur with higher frequency when a larger fraction of the nodes in a tree are caused by reassortment events rather than first-infection events.

Figure 8 shows the probability of inferring the correct number of reassortment events using two popu-

lar inference methods and an omniscient approach that counts the minimum number of remove-and-rejoin moves, i.e.,, the observed minimum, based on 100 simulated data sets for each condition. In the case where reassortments only occur in one segment, the omniscient approach is always correct, by definition, but both the coalescent-based [24] and topology-based [19] approaches tend to get the right number of reassortment events about 15%–45% of the time. However, when both segments are allowed to reassort, the naive method also tends to get the number of reassortment events wrong, because one of the segment trees is no longer guaranteed to have the same structure as the first-infection tree. We further investigated the deviation in `CoalRe` estimations of reassortment number and the reassortment timing across all the posterior ARGs in Fig. S6 and Fig. S7, respectively.

Discussion

The statistical situations that we laid out are all very challenging, if not entirely impossible to address starting from the coalescent framework that is popular in Bayesian phylogenetic applications. There are three problems that could debatably be called 'fundamental' in building a method of reassortment inference on the coalescent. First, the sampling rate in the coalescent is not defined as a meaningful population parameter. The coalescent conditions on the sample and works backward in time, such that the sampling rate (or probability in a batch sample) is not meaningfully defined. Even if the sampling rate can be inferred post-hoc (e.g., by computing the implicit rate after-the-fact) given the inferred population size dynamics [36], it would still be insufficient for adjusting for unsampled reassortment events, because the coalescent is, by definition, silent on the population history where unsampled reassortment events would most likely reside. Second, the coalescent has no obvious way of incorporating reassortment events in the context of the coalescent process itself. Reassortment is a coupled death-birth event that generates a node in a segment tree without changing the population size. Part of the reason that the coalescent is a popular formalism for Bayesian time-scaled phylogenetics is that it provides a theoretical basis for making, otherwise weakly identifiable, evolutionary rate models work. However, in the context of reassortment, some or even most nodes might have nothing to do with the population process assumed by the coalescent, and, therefore, are either wrongly included in the coalescence prior or are left with an implicitly ill-defined prior. Further, the coalescent with reassortment will, by necessity, not allow trifurcations or a single node to be both a reassortment and coalescent event, which prevents such a method from addressing two of the four situations that we raised. Third, reassortment in a coalescent framework only makes sense when paired with a corresponding coalescent event (in reverse time, a reassortment node creates a new lineage in an ARG that will eventually need to coalesce), so both the reassortment rate and the coalescent process itself become uninterpretable.

This is because the coalescent rate is, in most applications, computed as a function of the number of extant lineages and the effective population size. However, both first-infection and reassortment can result in a branching node in a tree, in which the former indeed increases population size, but the latter, formed by a coupled death-birth of an individual, does not change the population size. This convolution of the coalescent event and the reassortment event makes the interpretation of their corresponding rates inherently confusing.

We argued that a reliable method for inferring reassortment both in the number and timing of events needs to address both the fundamental situation with under-counting and the specific statistical situations that we raised. To our knowledge, no existing software or proposed method meets these criteria. We see three practical ways that already-published methods could be adapted to address some or all of these situations.

We demonstrated that simply conducting remove-and-rejoin moves to sampled segment trees is not a viable way to count the number of reassortment events that actually impacted the segment trees' structure, unless the first-infection tree is known. Existing topology-based methods (like those implemented in `TreeKnit`) could be adapted to search the space of remove-and-rejoin moves for a set of segment trees given a first-infection tree. In this context the first-infection tree is actually just a nuisance variable that we would need to integrate out, meaning we would need either an explicit or implicit probability mass function over the space of first-infection trees. Simulation methods could be used to generate first-infection trees from some prior distribution of model parameters, and existing remove-and-rejoin algorithms could be applied to all first-infection tree-segment tree pairs. Effectively this would both correct for the undercounting issue and would allow integration of prior knowledge in terms of the simulation model assumptions and the prior on the model parameters. Methods that jointly search both the tree space and model parameter space provide not only a more complete reporting of uncertainty, but also direct estimates of the timing of reassortment events. However, to our knowledge, existing methods are not able to address the situations we raise here.

We see several aspects of existing coalescent methods (such as `CoalRe`) that would need to be modified. First, the first-infection tree would have to be explicitly identified and manipulated as a fully or partially latent data structure. That is, ARG-based methods would not work as a first-infection tree cannot be meaningfully extracted from an ARG. However, a first-infection tree can be proposed by a well-defined population process. Given a first-infection tree, one then only needs the probability of remove-and-rejoin moves with specified reassortment times to link to the segment trees. (Mathematical details are provided in the Supplementary Materials.)

Our critique in this paper could be summarized as "to understand reassortment, one must place it in the context of a full population process." The problem is that there is no currently known method for computing the likelihood of either a sequence alignment or a set of segment trees given the population models that

we talk about in this paper. One potential way to deal with this is to treat the mapping between a population model and observable data through the lens of likelihood-free methods, such as Approximate Bayesian Computation (ABC) [37]. These methods work by defining a metric space based on some set of statistics computed on the observed data and data simulated from the population process of interest. Recently Voznica *et al.* [38] proposed a deep-learning approach along these lines where, rather than picking a set of statistics by hand, they allow a neural network to approximate a mapping between model parameters and simulated tree structures for determining the population model structure, estimating the parameters, and exploring the uncertainties. A similar approach might prove fruitful for fitting complex models with genomic rearrangements such as reassortment.

A final consideration for the use of population models, as both explicit theories and data analytic tools, is that a well-defined population model can serve as a common domain for scientific hypotheses and the statistical tests needed to probe those hypotheses. Consider that genomic reassortment can occasionally produce highly-fit viral strains that then rapidly spread in a population. It is easy to imagine how one might integrate this question into a population model, e.g., having a neutral type and an advantageous type (in terms of increased transmission rate) of reassortment. If we were able to fit that model to data, we could characterize a direct test of our scientific question as a clearly defined statistical test of the model (e.g., whether or not one transmission rate is larger than the other). However, without an explicit population model, any test of increased fitness will never be clean in that it will also be a test of all the auxiliary hypotheses that went into inferring the fact and timing of reassortment events. Depending on the methods used to infer reassortment, those assumptions could range from innocuous to objectively wrong (e.g., neutral evolution in constant population size). If we can never be certain that we are rejecting a hypothesis due to its wrongness, rather than the wrongness of some necessary but inconvenient auxiliary assumption made along the way, then our ability to test scientific hypotheses is compromised.

Inferring the existence and timing of macroscopic genomic rearrangements is exceptionally hard. Designing methods which can directly address the situations we laid out in this paper will lead to a clearer understanding of how to think about reassortment as a source of genetic diversity and, hopefully, improved methods.

Acknowledgements

This work was supported by grants from the U.S. National Institutes of Health (Grants #R01AI167048 to CMP, #R01AI087520 to TL, #R01AI143852 to AAK), and by the Joint U.S. National Science Foundation/National Institutes of Health Interface Program (Grant #1761603 to AAK).

Literature Cited

[1] Ulrich Desselberger, Katsuhisa Nakajima, Paul Alfino, Finn Skou Pedersen, William A Haseltine, Claude Hannoun, and Peter Palese. Biochemical evidence that "new" influenza virus strains in nature may arise by recombination (reassortment). *Proceedings of the National Academy of Sciences*, 75(7):3341–3345, 1978.

[2] Robert G Webster, William J Bean, Owen T Gorman, Thomas M Chambers, and Yoshihiro Kawaoka. Evolution and ecology of influenza a viruses. *Microbiological reviews*, 56(1):152–179, 1992.

[3] Etienne Simon-Loriere and Edward C Holmes. Why do rna viruses recombine? *Nature Reviews Microbiology*, 9(8):617–626, 2011.

[4] Catherine A Macken, Richard J Webby, and William J Bruno. Genotype turnover by reassortment of replication complex genes from avian influenza a virus. *Journal of General Virology*, 87(10):2803–2815, 2006.

[5] Andrew Rambaut, Oliver G Pybus, Martha I Nelson, Cecile Viboud, Jeffery K Taubenberger, and Edward C Holmes. The genomic and epidemiological dynamics of human influenza a virus. *Nature*, 453(7195):615–619, 2008.

[6] Roger 1 Glass, Jon Gentsch, and Jean C Smith. Rotavirus vaccines: success by reassortment? *Science*, 265(5177):1389–1391, 1994.

[7] Sarah M McDonald, Jelle Matthijnssens, John K McAllen, Erin Hine, Larry Overton, Shiliang Wang, Philippe Lemey, Mark Zeller, Marc Van Ranst, David J Spiro, et al. Evolutionary dynamics of human rotaviruses: balancing reassortment with preferred genome constellations. *PLoS Pathogens*, 5(10):e1000634, 2009.

[8] Barry J Beaty, Daniel R Sundin, Laura J Chandler, and David HL Bishop. Evolution of bunyaviruses by genome reassortment in dually infected mosquitoes (aedes triseriatus). *Science*, 230(4725):548–550, 1985.

[9] Thomas Briese, Charles H Calisher, and Stephen Higgs. Viruses of the family bunyaviridae: are all available isolates reassortants? *Virology*, 446(1-2):207–216, 2013.

[10] Richard M Elliott. Orthobunyaviruses: recent genetic and structural insights. *Nature Reviews Microbiology*, 12(10):673–685, 2014.

[11] T Yanase, T Kato, M Yamakawa, K Takayoshi, K Nakamura, T Kokuba, and T Tsuda. Genetic characterization of batai virus indicates a genomic reassortment between orthobunyaviruses in nature. *Archives of Virology*, 151:2253–2260, 2006.

[12] Thomas Briese, Brian Bird, Vishal Kapoor, Stuart T Nichol, and W Ian Lipkin. Batai and ngari viruses: M segment reassortment and association with severe febrile disease outbreaks in east africa. *Journal of Virology*, 80(11):5627–5630, 2006.

[13] David M Morens and Anthony S Fauci. Emerging infectious diseases: threats to human health and global stability. *PLoS Pathogens*, 9(7):e1003467, 2013.

[14] Stephen S Morse. *Factors in the emergence of infectious diseases*. Springer, 2001.

[15] Kevin C. Nixon. Phylogeny. In Simon A Levin, editor, *Encyclopedia of Biodiversity (Second Edition)*, pages 16–23. Academic Press, Waltham, second edition edition, 2001.

[16] Xavier Didelot, Christophe Fraser, Jennifer Gardy, and Caroline Colijn. Genomic infectious disease epidemiology in partially sampled and ongoing outbreaks. *Molecular biology and evolution*, 34(4):997–1007, 2017.

[17] Niranjan Nagarajan and Carl Kingsford. Giraf: robust, computational identification of influenza reassortments via graph mining. *Nucleic Acids Research*, 39(6):e34, 2010.

[18] Victoria Svinti, James A Cotton, and James O McInerney. New approaches for unravelling reassortment pathways. *BMC Evolutionary Biology*, 13:1–14, 2013.

[19] Pierre Barrat-Charlaix, Timothy G Vaughan, and Richard A Neher. Treeknit: Inferring ancestral reassortment graphs of influenza viruses. *PLoS Computational Biology*, 18(8):e1010394, 2022.

[20] C Randal Linder, Bernard ME Moret, Luay Nakhleh, and Tandy Warnow. Network (reticulate) evolution: biology, models, and algorithms. In *The Ninth Pacific Symposium on Biocomputing (PSB)*, 2004.

[21] J Maynard Smith and Noel H Smith. Detecting recombination from gene trees. *Molecular Biology and Evolution*, 15(5):590–599, 1998.

[22] Edward J Feil, Edward C Holmes, Debra E Bessen, Man-Suen Chan, Nicholas PJ Day, Mark C Enright, Richard Goldstein, Derek W Hood, Awdhesh Kalia, Catrin E Moore, et al. Recombination within natural populations of pathogenic bacteria: short-term empirical estimates and long-term phylogenetic consequences. *Proceedings of the National Academy of Sciences*, 98(1):182–187, 2001.

[23] Lena Collienne, Chris Whidden, and Alex Gavryushkin. Ranked subtree prune and regraft. *bioRxiv*, pages 2023–05, 2023.

[24] Nicola F. Müller, Ugnė Stolz, Gytis Dudas, Tanja Stadler, and Timothy G. Vaughan. Bayesian inference of reassortment networks reveals fitness benefits of reassortment in human influenza viruses. *Proceedings of the National Academy of Sciences*, 117(29):17104–17111, 2020.

[25] Ugnė Stolz, Tanja Stadler, Nicola F Müller, and Timothy G Vaughan. Joint inference of migration and reassortment patterns for viruses with segmented genomes. *Molecular Biology and Evolution*, 39(1):msab342, 2022.

[26] J. F. C. Kingman. The coalescent. *Stochastic Processes and their Applications*, 13(3):235–248, 1982.

[27] John Wakeley. *Coalescent Theory, an Introduction*. Roberts and Company, Greenwood Village, CO, 2005.

[28] James H Degnan and Noah A Rosenberg. Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends in Ecology & Evolution*, 24(6):332–340, 2009.

[29] Joseph Heled and Alexei J Drummond. Bayesian inference of species trees from multilocus data. *Molecular Biology and Evolution*, 27(3):570–580, 2009.

[30] Aaron A. King, Qianying Lin, and Edward L. Ionides. Markov genealogy processes. *Theoretical Population Biology*, 143:77–91, 2022.

[31] Willy Feller. Die grundlagen der volterraschen theorie des kampfes ums dasein in wahrscheinlichkeitstheoretischer behandlung. *Acta Biotheoretica*, 5(1):11–40, 1939.

[32] David G Kendall. On the generalized" birth-and-death" process. *The annals of mathematical statistics*, 19(1):1–15, 1948.

[33] Bui Quang Minh, Heiko A Schmidt, Olga Chernomor, Dominik Schrempf, Michael D Woodhams, Arndt Von Haeseler, and Robert Lanfear. Iq-tree 2: new models and efficient methods for phylogenetic inference in the genomic era. *Molecular Biology and Evolution*, 37(5):1530–1534, 2020.

[34] Remco Bouckaert, Timothy G Vaughan, Joëlle Barido-Sottani, Sebastián Duchêne, Mathieu Fourment, Alexandra Gavryushkina, Joseph Heled, Graham Jones, Denise Kühnert, Nicola De Maio, et al. Beast 2.5: An advanced software platform for bayesian evolutionary analysis. *PLoS Computational Biology*, 15(4):e1006650, 2019.

[35] David F Robinson. Comparison of labeled trees with valency three. *Journal of combinatorial theory, Series B*, 11(2):105–119, 1971.

[36] Kris V Parag, Louis du Plessis, and Oliver G Pybus. Jointly inferring the dynamics of population size and sampling intensity from molecular sequences. *Molecular biology and evolution*, 37(8):2414–2429, 2020.

[37] Simon Tavaré, David J Balding, Robert C Griffiths, and Peter Donnelly. Inferring coalescence times from dna sequence data. *Genetics*, 145(2):505–518, 1997.

[38] Jakub Voznica, Anna Zhukova, Veronika Boskova, E Saulnier, F Lemoine, Mathieu Moslonka-

Lefebvre, and Olivier Gascuel. Deep learning from phylogenies to uncover the epidemiological dynamics of outbreaks. *Nature Communications*, 13(1):3896, 2022.

[39] Joseph Felsenstein. Evolutionary trees from dna sequences: a maximum likelihood approach. *Journal of molecular evolution*, 17:368–376, 1981.

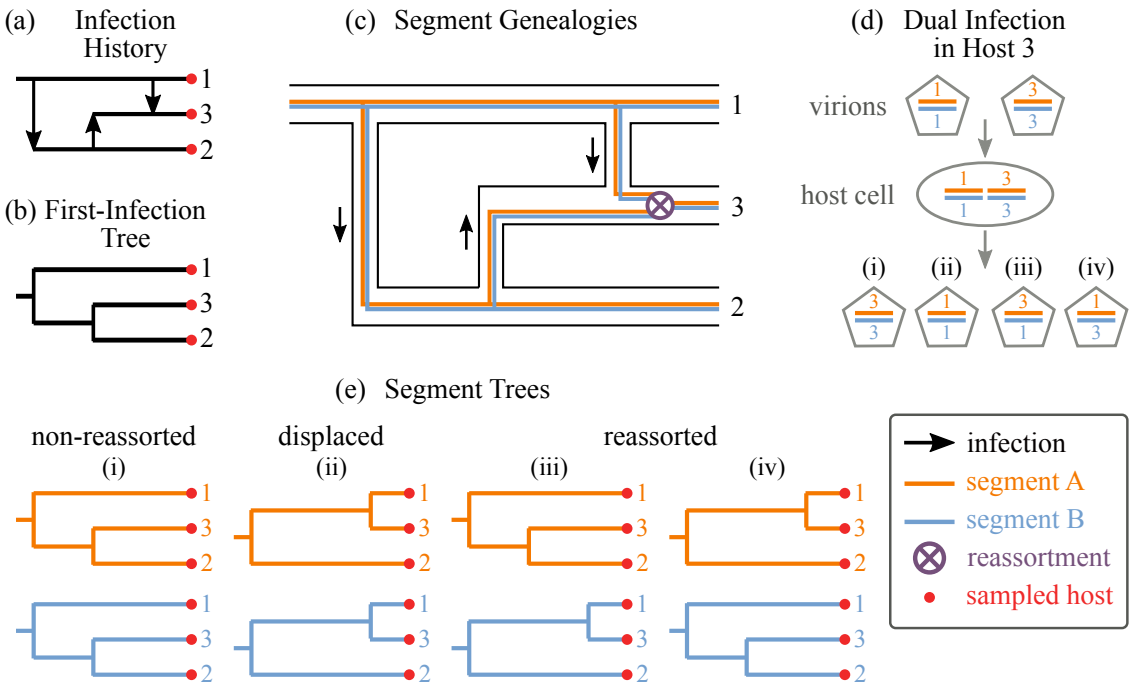


Figure 1: **Dual infection can lead to reassorted segment trees.** Time flows from left to right in panels a, b, c, and e. (a) The *infection history* encompasses all initial and subsequent infection events. In this small example, Host 1 infects Host 2, who then infects Host 3, and later Host 1 infects Host 3. Host 3 is then dually infected. (b) The *first-infection tree* includes initial infections of each host, but not subsequent infections or the identities of who infected whom. (c) The genealogy of each of the two viral genomic segments can be traced within the infection history. (d) Virions from the different infections of Host 3 (its pre-existing infection from Host 2, and its subsequent infection from Host 1) can reassort upon dual infection of a cell within Host 3. (e) Depending on which virion is sampled from Host 3, the *segment trees* may show discordance with one another and/or with the first-infection tree. In outcome (i) the host maintains the viral genotype from its first infection, in (ii) the genotype from the second infection displaces the first, and in (iii) and (iv) a new reassorted viral genotype is generated.

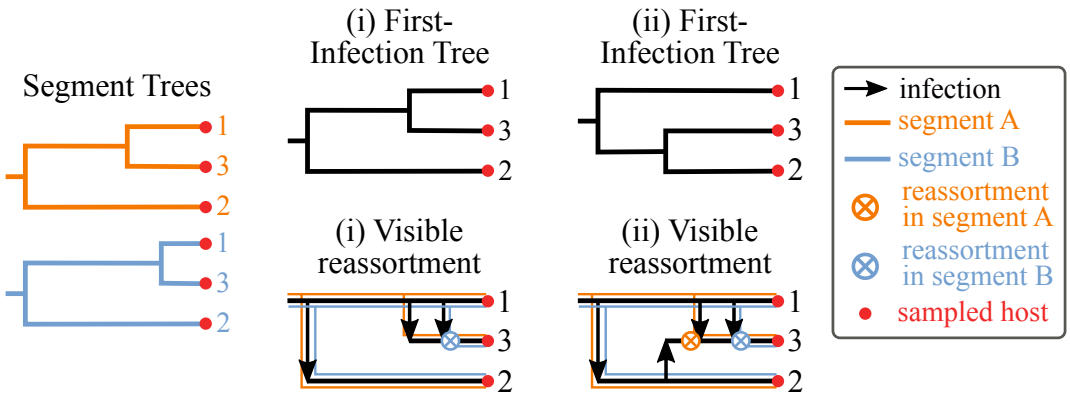


Figure 2: **The same segment trees with different infection histories require different reassortment explanations.** Suppose that two segment trees are sampled. In the special case that (i) the first-infection tree is equal to one of the segment trees, a single reassortment event is visible in the data. If (ii) the first-infection tree is different than either segment tree, two reassortment events are visible. Put another way, in (i) only reassortment in segment B is needed to explain the difference from the first-infection tree, while in (ii) reassortment in each segment is needed to explain why they both differ from the first-infection tree.

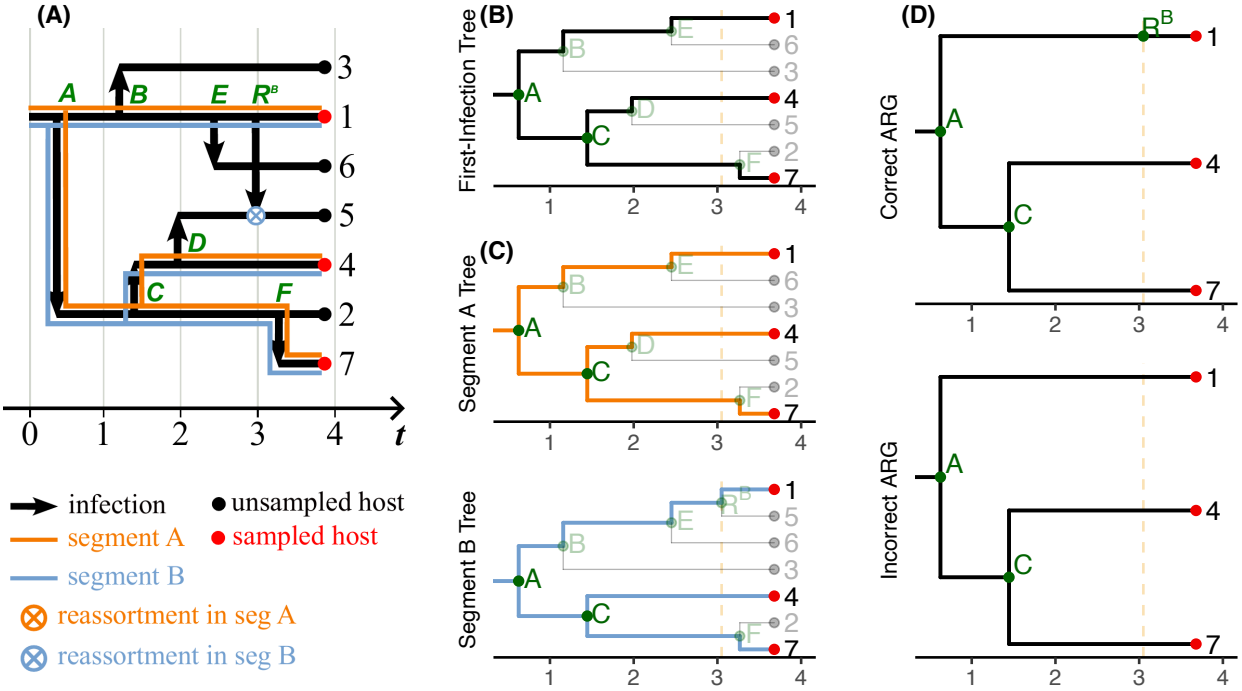


Figure 3: **Invisible reassortment.** This figure is an example of an *invisible* type of reassortment event that arises when a reassortment event either occurs in an unsampled part of the population or is replaced by a later reassortment event in the sampled part of the population. Trees are measured in arbitrary time units increasing from left to right. The full population history (A), first-infection tree (B), segment trees (C), and ARGs (D) are shown. In (D) the black dotted lines indicate an actual or inferred reassortment in the population and the orange dashed line indicates the time of that reassortment event.

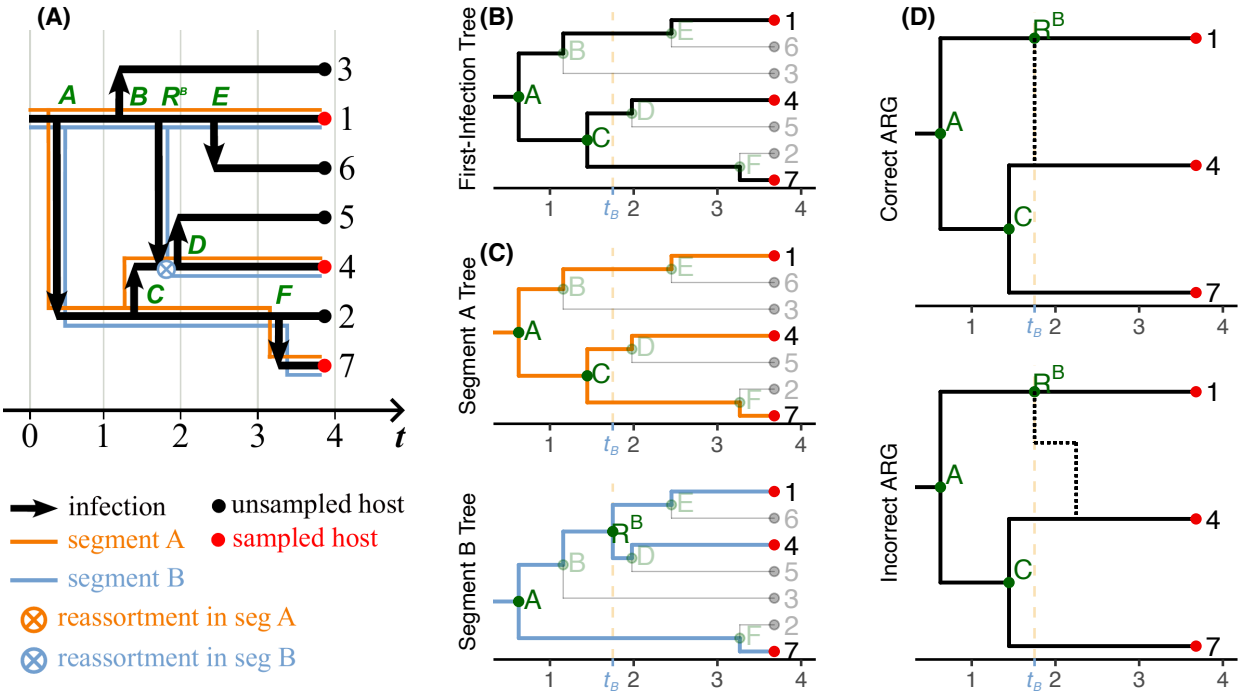


Figure 4: **Inaccurate reassortment.** This figure is an example of an *inaccurate* reassortment event that arises when both children of a reassortment event are sampled creating a node in the segment tree that was caused by reassortment and not a new infection. The full population history (A), first-infection tree (B), segment trees (C), and ARGs (D) are shown. In (D) the black dotted lines indicate an actual or inferred reassortment in the population and the orange dashed line indicates the time of that reassortment event.

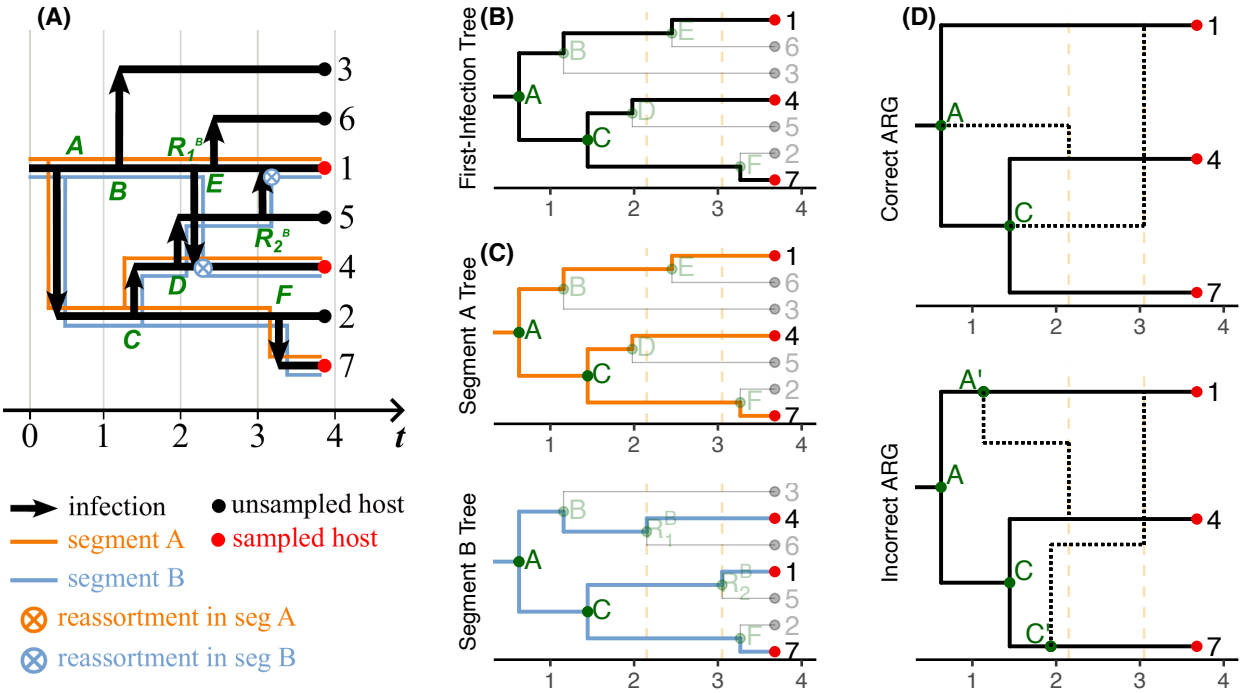


Figure 5: **Reversed reassortment.** This figure is an example of an *reversed* reassortment event that arises when only one child of a reassortment event is sampled creating the need for a trifrucating node in the ARG. The full population history (A), first-infection tree (B), segment trees (C), and ARGs (D) are shown. In (D) the black dotted lines indicate an actual or inferred reassortment in the population and the orange dashed line indicates the time of that reassortment event.

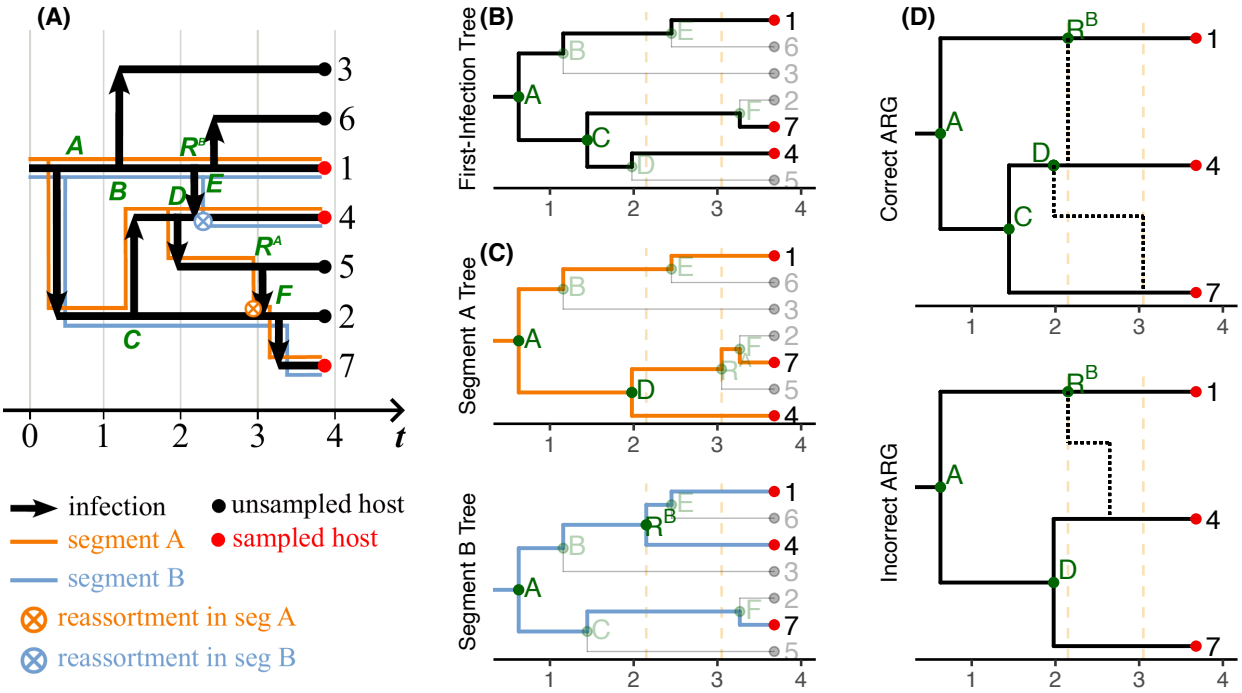


Figure 6: **Obscured reassortments.** This figure is an example of an *obscured* reassortment event that arises when both sampled segment trees contain reassortment events such that the first-infection tree no longer has the same structure as one of the segment trees. The full population history (A), first-infection tree (B), segment trees (C), and ARGs (D) are shown. In (D) the black dotted lines indicate an actual or inferred reassortment in the population and the orange dashed line indicates the time of that reassortment event.

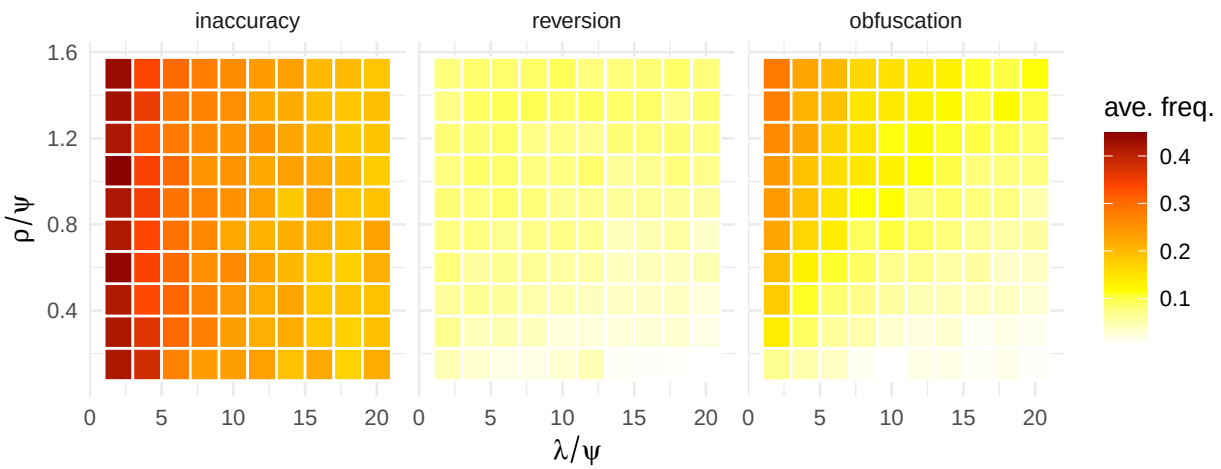


Figure 7: **The frequency of the *inaccuracy*, *reversion* and the *obfuscation* situations.** Starting from population size $N = 100$, the first infection rate is fixed at $\lambda = 5$ and the effective removal rate is $\mu + \psi = 5$, thus the population is asymptotically constant. The total reassortment rate ρ is the sum of the rates for each segment and the joint reassortment rate: $\rho = \rho_A + \rho_B + \rho_{AB}$, where $\rho_A : \rho_B = 3 : 2$ and $\rho_{AB} = 0$. Samples are taken sequentially at a constant rate ψ until 50 individuals in the population are sampled.

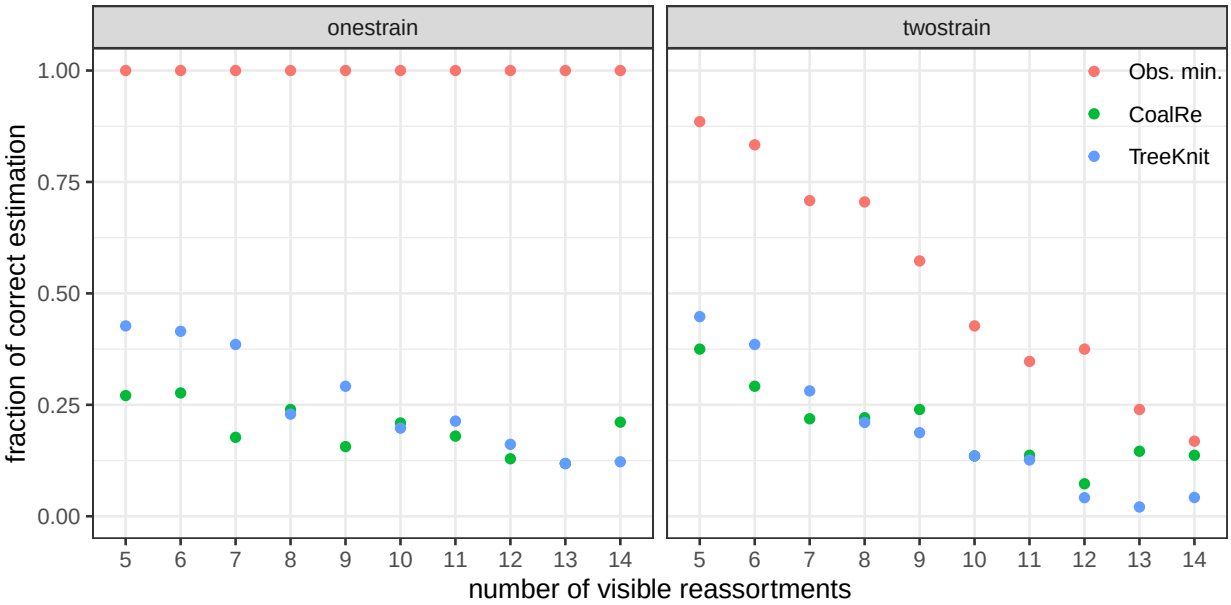


Figure 8: **Fraction of correctly estimating the number of visible reassortments**. In an asymptotically constant population, with $\lambda = 5$, $\mu = 4.75$, $\psi = .25$, and $\rho = .2$, we simulated 100 trees with 50 samples for each of the number of visible reassortments, ranging from 5 to 14. For each tree, we computed the observed minimum number of reassortments and estimated the number of reassortments by the inferred MCC ARG from `CoalRe` and the ARG from `TreeKnit`, such that we computed the fraction of 100 simulated trees that estimate the number of visible reassortments correctly. Two reassortment schemes are discussed: (1) only one segment reassorts, $\rho_A = .2$ and $\rho_B = \rho_{AB} = 0$, and (2) both segments reassort but not simultaneously, $\rho_A = .12$, $\rho_B = .08$, and $\rho_{AB} = 0$.

Modeling the evolution of segment trees reveals deficiencies in current inferential methods for genomic reassortment

Qianying Lin[1,†], Emma E. Goldberg[1], Thomas Leitner[1], Carmen Molina-París[1], Aaron A. King[2,3], and Ethan O. Romero-Severson[1,†]

[1]Theoretical Biology and Biophysics, Los Alamos National Laboratory, Los Alamos NM, USA

[2]Center for the Study of Complex Systems and Departments of Ecology & Evolutionary Biology and Mathematics, University of Michigan, Ann Arbor MI, USA

[3]Santa Fe Institute, 1399 Hyde Park Road, Santa Fe NM, USA

[†]Correspondence: *qianying@lanl.gov*, *eoromero@lanl.gov*

September 20, 2023

Supplementary Materials

Markov Genealogy Process with reassortments

Two essential elements for a Markov Genealogy Process with Reassortment are: (a) a population process $\mathscr{X}_t$ that describes the population dynamics, and (b) a segmented genealogy process $\mathscr{G}_t = (\mathscr{F}, (\mathscr{T}^s)_{s=1}^S)$ that incorporates the viral evolution along with the population dynamics, where S is the fixed number of segments. For a minimal model structure (i.e., a linear-birth-death-sample model) for a two-segment virus, we can define: (a) the population process as $\mathscr{X} = (I, G, M^A, M^A)$, where I is the infected population, G is the sample size, and M^A and M^B indicate the cumulative number of reassortments occurring in segment A and B, respectively; and (b) the genealogy process $\mathscr{G} = (\mathscr{F}, \mathscr{T}^A, \mathscr{T}^B)$, where $\mathscr{F}$ denotes the first-infection tree, and $\mathscr{T}^A$ and $\mathscr{T}^B$ denote the segment A and B trees, respectively. Table S1 summarizes how the four most important events, i.e., first infection, recovery, sampling, and reassortment, change the dynamics of the population $\mathscr{X}$ and the genealogies $\mathscr{G}$. Theoretical details can be found in [30].

Dismiss invisible reassortments

In a simulated tree, an orange node marks a reassortment that has occurred in the other segment tree and a gold node denotes a reassortment that has occurred in the respective segment tree. Figure S3 left panels show a raw first-infection tree and the corresponding raw segment tree, before dismissing invisible reassortment nodes. The first-infection tree only contains orange reassortment nodes because no reassortment is allowed in the first-infection tree, while the segment tree only contains gold reassortment nodes because only this one segment reassorts in this example. We highlight the discording branching nodes in light green and purple shades, where purple shades denote the reversed branching nodes. Reassortments can become invisible in the sampled tree due to later reassortments along the same branch or on the same clade. For instance, in the first-infection tree at the left panel of Fig. S3, reassortment node No. 75 is invisible due to No. 117, and No. 15 is invisible due to No. 58 and No. 117. By dismissing those invisible reassortment nodes in both trees, we can have the minimum number of reassortments that correspond to the observed structural difference between the first-infection tree and the segment tree, as shown in the Fig. S3 right panels.

The visible reassortment events and the remove-rejoin table

Figure S4 shows how one performs the 'remove-and-rejoin' method and counts the minimum number of reassortments given the first-infection tree and segment trees. After dismissing all invisible reassortment nodes, one can conduct the 'remove-and-rejoin' moves by comparing the first-infection tree and a segment tree. For example, removing the clade associated with reassortment node a_1 also removes branching node E, while reattaching it back to the clade of branching node B creates a new branching node R_1^A, and eventually

reduces the structural differences between the first-infection tree and the segment A tree. By repeating these remove-and-rejoin moves between the first-infection tree and each of the segment trees, one can write down two 'remove-and-rejoin' tables. Note that, node c_1 implies that the corresponding reassortment occurs in both segments A and B simultaneously. This reassortment only causes a difference in branch length between the first-infection tree and the segment trees, while excluding which may result in biased estimation in mutation rates for reconstructed phylogenetics. Combining two tables and eliminating the duplicates gives us minimal remove-and-rejoin moves and the visible number of reassortments. With the help of the remove-rejoin table, one can produce the correct ARG.

When one lacks knowledge of the first-infection tree, one can conduct the identical procedure to produce the 'observed' remove-rejoin table, after faking a first-infection tree and a 'pure' segment tree. One can choose one segment tree at random to become the 'virtual' first-infection tree, by changing all the gold nodes to orange nodes and changing all orange nodes to gold nodes to make a 'pure' segment tree. Fig. S5 illustrates this procedure.

The number and frequency of three situations among visible reassortments

Using the simulated segment trees, one can easily identify the inaccuracy situation by counting the number of reassortment nodes that are aligned with its ancestral branching nodes (Table S1). For example, in either the segment A tree or segment B tree in Fig. S4, reassortment node c_1 is aligned with the new branching node R_1^C, indicating the reassortment time of c_1 is identical to the branching time of R_1^C. The reassortment node c_1 falls into the *inaccuracy* situation category.

Reversion indicates that a seemingly consistent branching node in both trees, i.e., same node label and branching time, actually has distinguished descendants in two trees. For instance, branching node D appears in both the first-infection tree and the segment A tree in Fig. S4, with different descendant tips. We then call node D a reversed node, and the count for the *reversion* situation between the first-infection tree and the segment A tree is one.

Obfuscation is the most abstract situation among the four proposed, however, the easiest to identify. One can produce the visible remove-rejoin table by comparing the first-infection tree and each of the two segment trees, as shown in Fig. S4. One then can produce the observed remove-rejoin table after the procedure proposed in Fig. S5. Extracting the true remove-rejoin table by the observed remove-rejoin table helps identify the degenerated moves.

The frequency of problematic situations can be computed by dividing the counts of the respective situation by the number of visible reassortments. For instance, in Fig. S4, the number of visible reassortments is

4, among which reassortment c_1 corresponds to the *inaccuracy* situation, reassortment node a_1 or a_2 corresponds to the *reversion* situation. We thus have the frequency of the *inaccuracy* situation is $1/4$ and that of the *reversion* is also $1/4$.

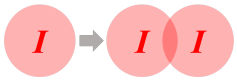
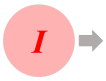
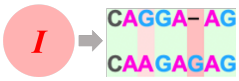
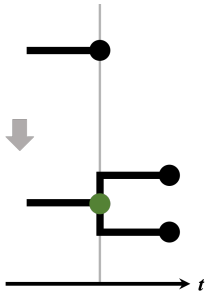
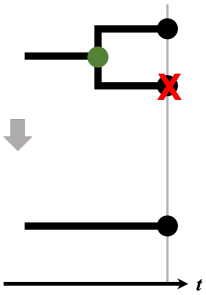
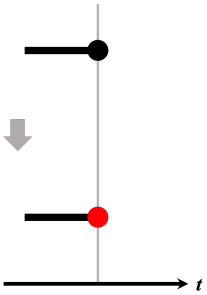
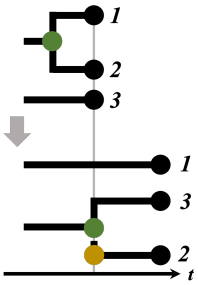
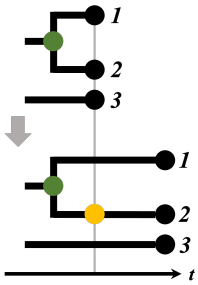
Adding the first-infection tree to the ARG likelihood function

For existing coalescent methods (such as `CoalRe`) to address the situations we identified in this paper, the first-infection tree would have to be incorporated in the likelihood function along with the segment trees. The classic Felsenstein likelihood [39] of the segment trees is then the last ingredient in the posterior distribution:

$$\mathbb{P}\left(\vec{\mathscr{T}}, \mathscr{F}, \theta_M, \vec{\varepsilon}, \vec{\rho} \mid \vec{A}\right) \propto \mathbb{P}\left(\vec{A} \mid \vec{\mathscr{T}}, \vec{\varepsilon}\right) \mathbb{P}\left(\vec{\mathscr{T}} \mid \mathscr{F}, \vec{\rho}, \theta_M\right) \mathbb{P}(\mathscr{F} \mid \theta_M)\, \mathbb{P}(\theta_M, \vec{\varepsilon}, \vec{\rho}).$$

Here, θ_M is the parameters for a given population model M; $\mathscr{F}$ is the first-infection tree induced by the given model M and parameters θ_M; elements in vectors $\vec{\mathscr{T}}$, $\vec{A}$, and $\vec{\varepsilon}$ represent the corresponding segment tree, alignment, and mutation model, for the respective segment, separately; and $\vec{\rho}$ are the reassortment rates for each and all combinations of segments. The expression on the right-hand side consists of the likelihood of segment trees $\mathbb{P}\left(\vec{A} \mid \vec{\mathscr{T}}, \vec{\varepsilon}\right)$, the segment trees prior given remove-and-rejoin tables from the first-infection tree $\mathbb{P}\left(\vec{\mathscr{T}} \mid \mathscr{F}, \vec{\rho}, \theta_M\right)$, the first-infection tree prior $\mathbb{P}(\mathscr{F} \mid \theta_M)$, and the joint parameter prior $\mathbb{P}(\theta_M, \vec{\varepsilon}, \vec{\rho})$. Specifically, by including both the topological difference and the difference in branch length between the first-infection tree and the segment trees in the remove-and-rejoin tables, that is in $\mathbb{P}\left(\vec{\mathscr{T}} \mid \mathscr{F}, \vec{\rho}, \theta_M\right)$, plays a role in inferring the mutation model $\vec{\varepsilon}$.

Table S1: **Summary of the population process, the genealogy process, and the important events.** The population process $\mathscr{X}$ is framed into a birth-death-sample model, represented by the number of infected population I, the cumulative count of samples G, and the cumulative counts (i.e., M^A and M^B) of reassortments of Segment A and B, respectively. The genealogy process $\mathscr{G}$ consists of three genealogies, $\mathscr{F}, \mathscr{T}^A, \mathscr{T}^B$. Four important events and their impact on $\mathscr{X}$ and $\mathscr{G}$ are summarized: (a) a first infection event increases I by one and introduces a new branch and a new lineage in all three genealogies in $\mathscr{G}$, at rate λ; (b) a recovery event reduces I by one and removes a tip and its corresponding branch from all genealogies, at rate μ; (c) a sampling event increases G but decreases I, and marks a tip as a sample then terminates it in all three genealogies, at rate ψ; and (d) a reassortment event in segment A increases the count M^A by one, then performs a death-birth operation, i.e., destroy and reattach a lineage, in only $\mathscr{T}^A$, marking the respective lineage by a gold node, while marking the same lineage in $\mathscr{F}$ and $\mathscr{T}^B$ by an orange node to demonstrate the correspondence, at rate ρ_A.

first infection	**recovery**	**sampling**	**reassortment**	
				
event rate (per lineage):				
λ	μ	ψ	ρ_A (per segment)	
population process: $\mathscr{X} = (I, G, M^A, M^B)$				
$I = I + 1$	$I = I - 1$	$G = G + 1, I = I - 1$	$M^A = M^A + 1$	
genealogy process: $\mathscr{G} = (\mathscr{F}, \mathscr{T}^A, \mathscr{T}^B)$				
in $\mathscr{F}, \mathscr{T}^A, \mathscr{T}^B$:	in $\mathscr{F}, \mathscr{T}^A, \mathscr{T}^B$:	in $\mathscr{F}, \mathscr{T}^A, \mathscr{T}^B$:	in $\mathscr{T}^A$:	in $\mathscr{F}, \mathscr{T}^B$:
				

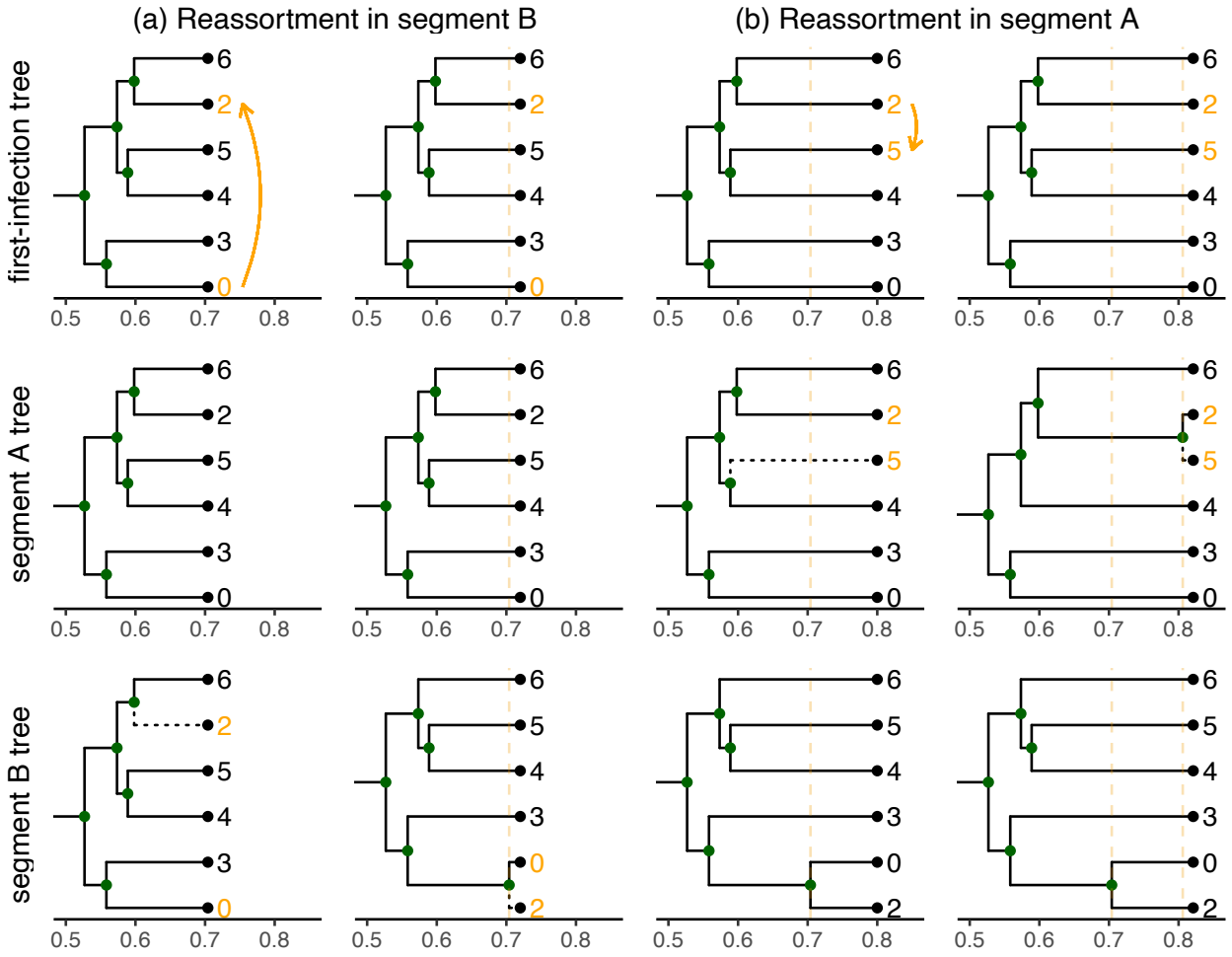


Figure S1: **Reassortment event in the genealogy process for a two-segment virus**. The top and bottom panels show how genealogy evolves for segments A and B at a reassortment event, respectively. The left panels represent the segmented genealogies of seven infected individuals in the population, right before a reassortment occurrence; During the reassortment event, host 0 serves as the donor and transmits the virus to host 2, the recipient, resulting in the swapping of genetic materials in segment B between hosts 0 and 2. In segment B genealogy, the reassortment event will be reflected by the death of host 2 followed by the rebirth of them in host 0 (middle panels). The right panels are the resulting segmented genealogies after the reassortment event. Infected individuals are represented by black dots, and the internal nodes in genealogy are in green; the vertical dashed line in orange indicates the reassortment event time.

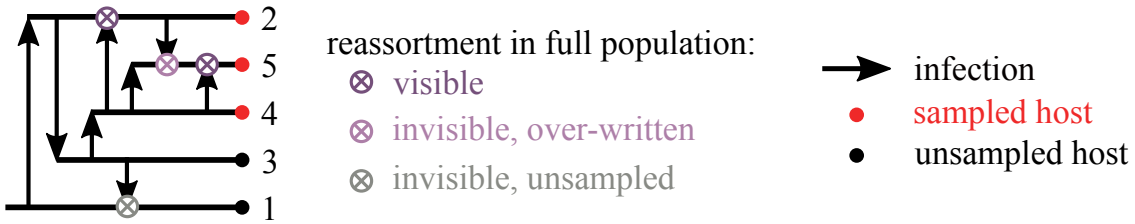


Figure S2: **Visible and invisible reassortments.** Of all the reassortment events in the full population (all four ⊗ shown), only some leave a visible imprint in the data. Reassortment events that occur in unsampled lineages are invisible, as are those that occur in sampled lineages but are over-written by another reassortment event later in the same lineage.

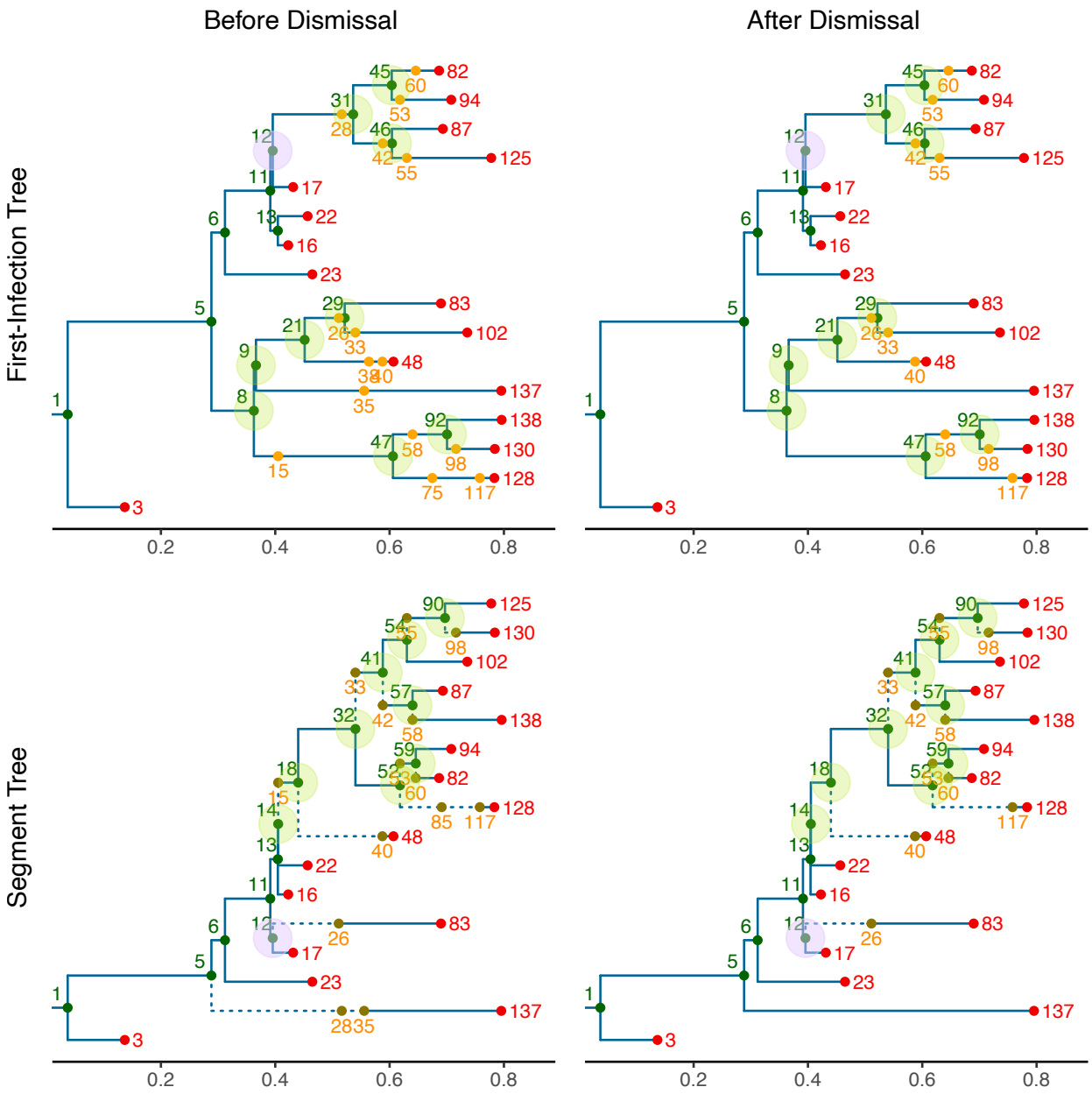


Figure S3: **Dismiss invisible reassortments.** This figure serves as a demonstrative example of the Dismissal Procedure. Our implementation records every node/tip and its label in the trees: shown in green, red, or orange/gold in the figure. Specifically, reassortment events are indicated by orange nodes in the first-infection trees and gold nodes in the segment tree that conducts the corresponding reassortments. Discording branching nodes are highlighted in green shape and the reversed branching nodes are in light purple shape. Before dismissal, in total 15 reassortment events are preserved in the sample trees, but some of them are not shared by both trees or in fact overwritten by the latter ones and will never be observed among samples. After dismissing these invisible reassortments, in total 10 reassortments were left, corresponding to 9 discording branching nodes and one reversed branching node between the first-infection tree and the segment tree. The remained reassortments are the minimum that is responsible for the discording and reversed branching numbers between trees.

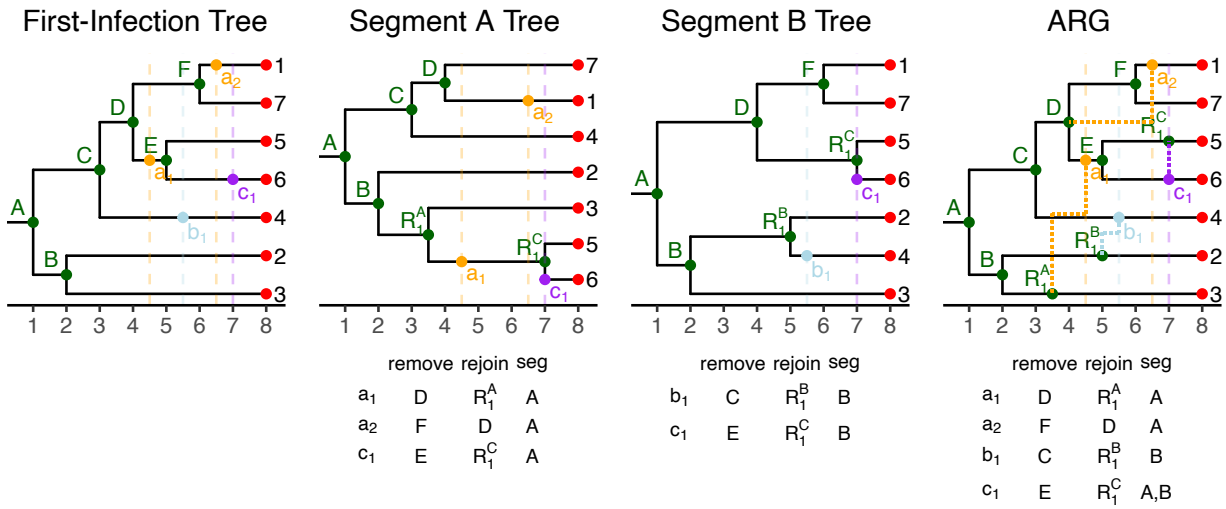


Figure S4: **The number of visible reassortments and the remove-and-rejoin tables.** By comparing the first-infection tree and each segment tree after dismissal, one can derive a remove-and-rejoin table for each segment tree. The union of all remove-rejoin tables reflects the true ARG and the number of visible reassortments. Nodes a_1 and a_2 indicate the location (i.e., the corresponding lineage and the reassortment time) of reassortment in segment A tree, b_1 in segment B tree, and c_1 in both.

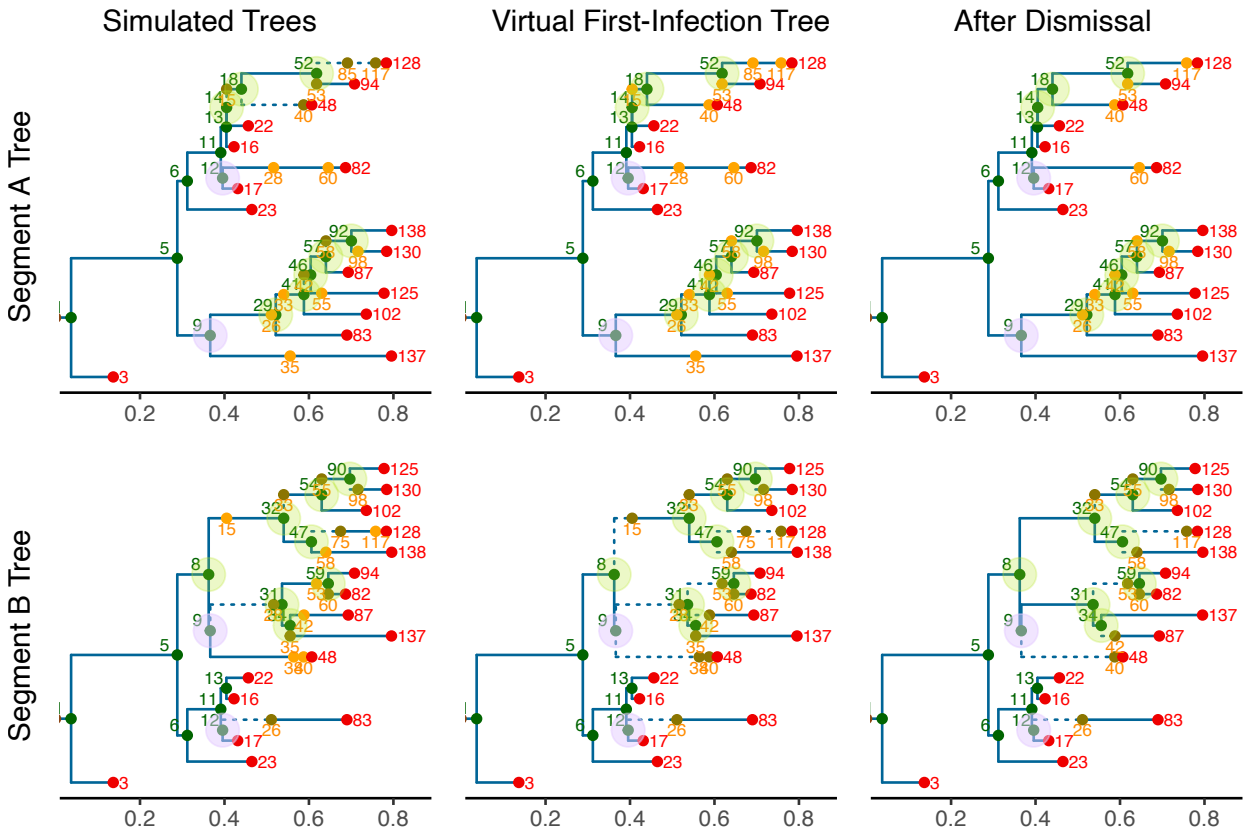


Figure S5: **Compute the number of visible reassortments given segment trees only.** Lacking the true first-infection tree, we select the segment A tree as the *virtual* first-infection tree. Given the *raw* simulated segment trees, all gold nodes in the segment A tree, indicating reassortments occurring in Segment A, are turned into orange, serving as markers for reassortment location. Correspondingly, all orange nodes are turned into gold in the segment B tree to produce a *pure* segment tree. Given the virtual first-infection tree (i.e., segment A tree) and the pure segment tree (i.e., segment B tree), one can perform the dismissal procedure (Figure S3) and compute the number of *observed* visible reassortments by performing remove-and-rejoin moves (Figure S4).

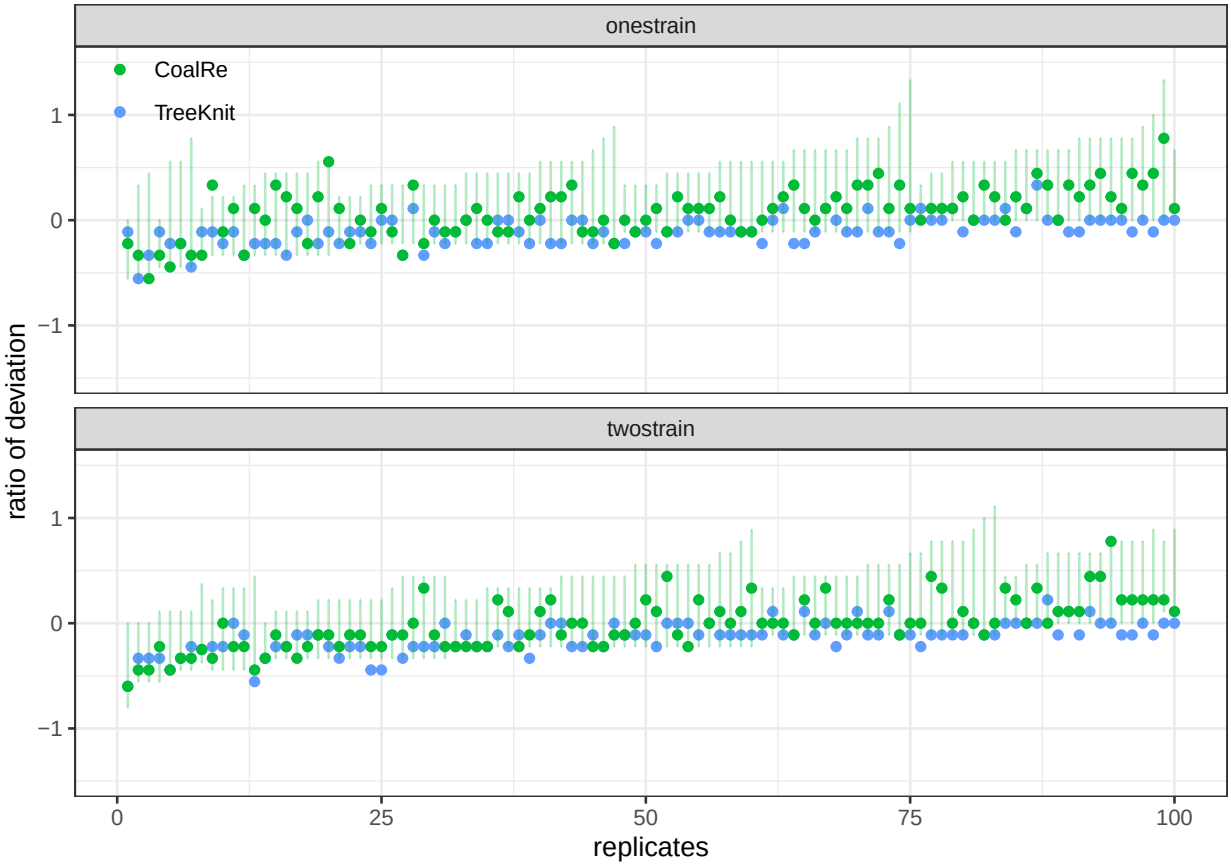


Figure S6: **Closer look at the estimated number of reassortments by `CoalRe` and `TreeKnit`.** In an asymptotically constant population, with $\lambda = 5$, $\mu = 4.75$, $\psi = .25$, and $\rho = .2$, we simulated 100 trees with 50 samples for each of the number of visible reassortments, ranging from 5 to 14. The number of visible reassortments is 9 for both panels. In `CoalRe`, 500 posterior ARGs are used to compute the maximum clade credibility tree (green dot) for each replicate, and the green bars indicate the 95% HPD among those realizations. Two reassortment schemes are discussed: (1) only one segment reassorts, $\rho_A = .2$ and $\rho_B = \rho_{AB} = 0$, and (2) both segments reassort but not simultaneously, $\rho_A = .12$, $\rho_B = .08$, and $\rho_{AB} = 0$.

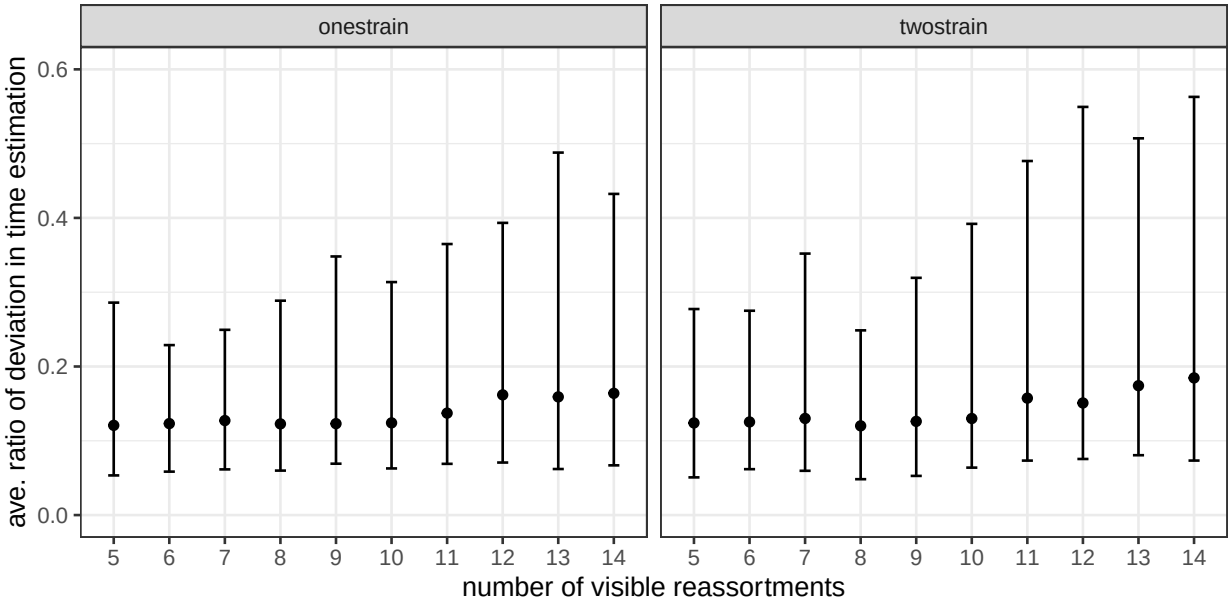


Figure S7: **Deviation in reassortment time estimations by `CoalRe`.** In an asymptotically constant population, with $\lambda = 5$, $\mu = 4.75$, $\psi = .25$, and $\rho = .2$, we simulated 100 trees with 50 samples for each of the number of visible reassortments, ranging from 5 to 14. For each simulated tree, `CoalRe` produced 500 posterior ARGs and some of them estimated the number of visible reassortments correctly. Among those getting the correct estimation, we computed the average RMSE of reassortment time between each of the posterior ARG by `CoalRe` and the simulated tree, normalized by the height of the simulated tree. We define this metric as the *average ratio of deviation in time estimation*. We show the 95% quantile and the median of this metric for each of the number of visible reassortments. Two reassortment schemes are discussed: (1) only one segment reassorts, $\rho_A = .2$ and $\rho_B = \rho_{AB} = 0$, and (2) both segments reassort but not simultaneously, $\rho_A = .12$, $\rho_B = .08$, and $\rho_{AB} = 0$.