Using multiple infection history to infer contact network structure

Christian Selinger[a,1] **and Samuel Alizon**[a]

[a]Laboratoire MIVEGEC (UMR CNRS 5290, IRD 224, UM), Montpellier, France

This manuscript was compiled on July 12, 2019

The structure of host interactions within a population shapes the spread of infectious diseases but contact patterns between hosts are difficult to access. We hypothesised that key properties of these patterns can be inferred by using multiple infections data from individual longitudinal follow-ups. To show this, we simulated multiple infections on a contact network in an unbiased way by implementing a non-Markovian extension of the Gillespie algorithm for a community of parasites spreading on this network. We then analysed the resulting individual infection time series in an original way by introducing the concept of 'infection barcodes' to represent the infection history in each host. We find that, depending on infection multiplicity and immunity assumptions, knowledge about the barcode topology makes it possible to recover key properties of the network topology and even of individual nodes. The combination of individual-based simulations and barcode analysis of infection histories opens promising perspectives for the study of infectious disease transmission networks.

network | epidemiology | coinfections | infection history | barcodes

Host populations are often assumed to be 'well-mixed' even though individuals tend to only interact with a small subset of the whole population and these contact patterns between individuals can dramatically affect the way epidemics spread (1–3). For instance, it was shown during the early phase of the HIV pandemics that the average number of sexual partners but also the variance in the number of partners both increase the basic reproduction number (R_0) of sexually transmitted infections (4). Since then, studies have identified how key parameters of the host contact network affect the risk of outbreak (5, 6) and the spread of an epidemic (7–12).

Measuring host contact networks is a lively research topic (13), which ranges from definition issues (defining a 'contact'), to assessing the appropriateness of various types of data. In some cases such as networks connecting farms together (14), interactions between individual units can be well approximated through shipping logs. Other settings, such as wild populations or human populations, are more challenging to analyze. For humans, the field has traditionally relied on self-reported data, but new insights have been provided by airline transportation registries (15) or cell phone data (16). More recently, parasite sequence data was used to infer network properties by analysing phylogenies of infections (17). Note that we use the term parasite in its evolutionary ecology meaning, which encompasses both micro- and macro-parasites. Importantly, this genetic data, which by definition pre-dates new outbreaks, can be used to make relevant predictions (18).

We hypothesize that key properties from host contact networks can be inferred from individual longitudinal data about infection status. Such longitudinal data is classically used in epidemiological studies to measure the odds that a specific event may occur (19), however they are rarely coupled to mathematical models of disease spread.

In our case, we also track multiple parasite strains or species in these individual histories. This is both feasible and highly relevant, assuming that parasites spread on the same network. Indeed, the increasing affordability and power of sequencing strengthens the assertion that most infections are genetically diverse (20). In fact, in the case of genital infections by human papillomaviruses (HPVs), not only do we know that coinfections, i.e. the simultaneous infection by multiple genotypes, are the rule rather than the exception, we also know that they strongly correlate with the number of lifetime sexual partners (21). Another classical example is the distribution of the number of macroparasites per host, which has been used to infer population structure (22). This makes multiple infections an ideal candidate to measure epidemic properties (23).

To show how individual infection histories for multiple parasite strains can inform us about the underlying transmission contact network, we conduct a simulation study, which requires alleviating two obstacles. First, we need to simulate multiple infections on a network. This in itself is not an easy task and few studies have attempted it (12, 24–29). Second, we need to extract information from longitudinal follow-up data. For this, we take advantage of recent development in stochastic epidemiological modelling and implement a non-Markovian version of the well-known Gillespie algorithm (30). This allows us to make sure a host's infection history is not lost every time it acquires a new strain. It also addresses statistical evidence that for many parasites infection duration does not follow exponential but heavy-tailed distributions (31–36). Finally, we need to compare infection histories, which we do by using barcode theory from computational topology (37).

We first illustrate the importance of infection histories

Significance Statement

The way hosts interact with each other is known to shape epidemics spread. However, these interactions are difficult to infer, especially in human populations. Using recent developments in stochastic epidemiological modeling and barcode theory, we show that the diversity of infections each host has undergone over time contains key information about contact network between hosts. This means that longitudinal follow-ups of some individuals in a population can tell us how hosts are in contact with each other. It can also inform us on how connected a particular individual is. This opens new possibilities regarding the use of genetic diversity of infectious diseases in epidemiology.

CS and SA conceived the study, CS generated and analyzed the data, CS and SA wrote the manuscript.

The authors declare no conflict of interest..

[1]To whom correspondence should be addressed. E-mail: christian.selinger@ird.fr

DRAFT

when modelling multiple infections. We find that accounting for the documented fact that recovery from infection becomes more likely with increasing age of infection (38–41) leads to higher prevalence peaks and earlier extinction times than in memory-less models. Then, we simulate individual multi-strain infection histories for epidemiological models with genotypes-specific immunity on random clustered networks to demonstrate the impact of network topology on infection diversity. We refer to 'genotypes' to describe the parasite diversity, but our results apply to different genotypes of the same species or of different species of parasites provided that they are spreading on the same network. Finally, we show that infection barcodes can inform us on the connectivity in the network in several ways. First, similarity matrices between individual infection histories highly correlate with the network adjacency matrices. Second, an individual host's network connectivity can be inferred from its connectivity in the infection barcode matrix.

Results

Memory in multiple infections. Under the assumption of neutrality and absence of interactions between genotypes, coinfections should intuitively follow the principle of 'first in, first out', i.e. the parasite genotype which is acquired first should also be cleared first from the host. To show that memory-less coinfection processes violate this principle, we simulate epidemics where we sequentially introduce four genotypes on random clustered networks and assume that full genotype-specific immunity is conferred once a host recovered from infection. The model is parameterized such that recovery from infection episodes follows exponential (memory-less) or Weibull (with memory) distribution at unit mean (Figure 1A). Corresponding hazard rates used as model inputs are constant or increase linearly with the age of infection respectively.

We find that the assumption of recovery with memory impacts epidemic outcomes with increased peaks, shorter duration and higher infection multiplicity when compared to the case of memory-less recovery (Figure 1B). Moreover, sequential introduction of different genotypes leads to further imbalance of infection and recovery events. Figure 1C shows the prevalence time series for 4 genotypes introduced one time unit apart from each other, genotype 1 being introduced first. Because we assume neutrality, their order of introduction should not matter. However, without memory (in red), the earlier a genotype is introduced during the epidemic, the higher prevalence because hosts are more likely to recover but also to create new infections. Those introduced later will initially be less likely to produce recovery or infection events, but lead to sustained epidemics later on, once most of the genotypes introduced first have been cleared and hosts become immune to them. Taking into account the age of infection for recovery (in blue) can alleviate this imbalance such that average epidemic curves and final epidemic sizes are similar between genotypes therefore resembling a scenario where all genotypes are introduced simultaneously (data not shown).

Network topology and multiple infections. To better understand how the topology of a contact network impacts multiple infection dynamics, we simulate SIR epidemics with simultaneous introduction of four infections by different genotypes on random clustered graphs with distinct summary statistics.

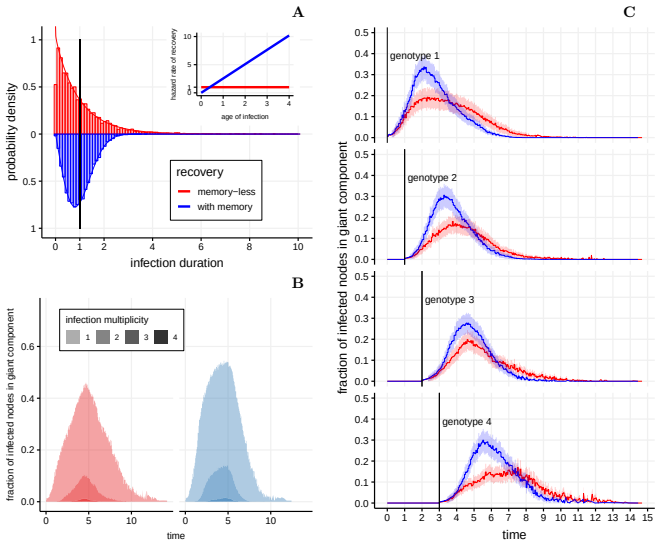

Fig. 1. Memory and multiple infections dynamics. Multiple infections are simulated on random clustered networks assuming recovery without (red) or with (blue) memory. A) Histograms of simulated individual infection durations and their parametric fit to exponential and Weibull distributions (smooth lines) as well as the corresponding hazard rates of recovery as a function of the age of infection used as model inputs. B) Total fraction of the population that is infected as well as the averaged multiplicity of infection over time. C) Averaged epidemic curves and 95% confidence intervals (shaded areas). Subpanels in C show the fraction of infected nodes of four different genotypes introduced sequentially into the host population. Parameter values are default (see Material and Methods) with 50 stochastic replicates per setting.

First, we consider random clustered graphs with low average but highly over-dispersed degree, and low clustering coefficient. These emulate contact networks with relatively low but variable number of contacts spread out evenly in the network (type 1 in Table 1). With a multiplicity of infection (MOI) of 1.5 and a diversity index of 3 at the peak of the epidemic (Figure 2), simulation results indicate that an average host is singly infected and that three out of the four genotypes circulate at equivalent prevalence. Next, we increase the clustering coefficient while keeping average node degree low, which results in less dispersed degree distributions emulating networks with highly clustered pockets of contacts but generally small number of contacts (type 2 in Table 1). Multiple infections on such networks have very similar MOI patterns but slightly increased and more sustained diversity of genotype prevalence (Figure 2). Finally, we consider networks with both high degree and high clustering coefficient resulting in higher degree assortativity such that the highest average MOI is approximately 2 during the epidemic peak (type 3 in Table 1). During the initial expansion phase of such epidemics, a typical host is coinfected, and genotype prevalence is more unevenly distributed due to high degree assortativity.

Table 1. Network summary statistics

Network type	degree	degree assortativity	degree dispersion	clustering coefficient
1	4.6	0.58	51.51	0.13
2	4.1	0.53	2.36	0.36
3	8.1	0.68	6.52	0.38

Quantities are averaged across 250 nodes and 50 stochastic replicates.

DRAFT

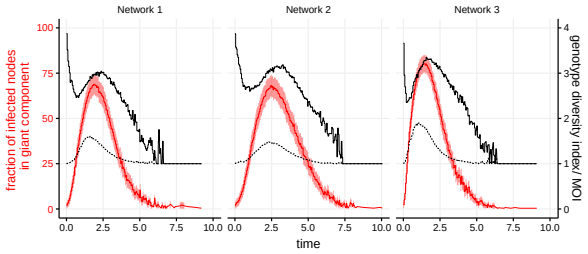

Fig. 2. Genotype diversity index and multiplicity of infection (MOI) for SIR epidemics with four genotypes on networks with distinct characteristics (average degree and clustering coefficient). For each panel, the red line shows the time-averaged fraction of nodes infected with at least one genotype (95% confidence intervals shaded red), the black solid line shows the averaged genotype diversity index at the population level, and the black dotted line shows the average multiplicity of infection of individual nodes. We simulated 50 replicates for each parameter setting with network parameter values according to Table 1.

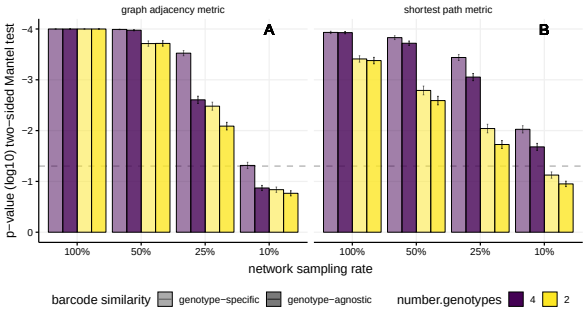

Fig. 3. Spatial correlations between an infection matrix based on barcode and A) the network's adjacency matrix or B) the network's shortest path matrix. We simulated SIR epidemics with 2 or 4 circulating genotypes using p-values of two-sided Mantel tests with 10^4 permutations. For each combination, we resampled 20 times randomly $100, 50, 25$ or 10% of infected nodes and report the average p-value and the 95% confidence interval.

Infection barcodes and network structure. To infer properties of the host contact network from individual infection histories, we describe each of these through an infection barcode, i.e. a set containing all intervals of infection episodes with different parasite genotypes. In order to measure similarity (or distance) between infection barcodes of any two hosts, we use the length and frequency of overlaps between infection episodes. As detailed in the Methods, we can consider overlaps for matching genotypes only or for any genotypes, the latter being less informative. This allows us to obtain barcode matrices that indicate the similarity (or distance) between any two nodes. These can then be compared to more classical matrices that capture global (shortest path between nodes) or local properties (binary adjacency) of contact networks restricted to infected individuals.

Simulation results with four different parasite genotypes show that, if the entire network of infected individuals is known, the barcode similarity matrix highly correlates with local properties of the contact network given by its adjacency matrix (Figure 3A). In more realistic situations, infected nodes would only be partially observed. Our results show that spatial correlations become less significant as the sampling rate of network nodes decreases but remains significant for sampling proportions of 25% or more. Given the importance of multiple infection histories in this approach, decreasing the number of circulating genotypes from 4 (violet) to 2 (yellow) also decreases inference power. We also investigate spatial correlations between the barcode similarity matrix and global properties such as the shortest path metric. We find that the correlations are highly significant even at a network sampling rate of 10% (Figure 3B), but suffer similar limitations if the number of genotypes circulating in the epidemic is reduced to 2.

Overall, we found that the barcode matrix built using the genotype-specific similarity index was more informative than that built on the distance function. Correlation coefficients obtained from the Mantel test are as high as 0.3 for comparisons of barcode similary matrices with shortest path matrices regardless of network sampling rates and number of genotypes (Figure S2). Alternatively, considering barcode distance matrices gives highly significant correlations with the graph adjacency metric, but no significant correlations with shortest path metric, even at high network sampling rates (Figure S1).

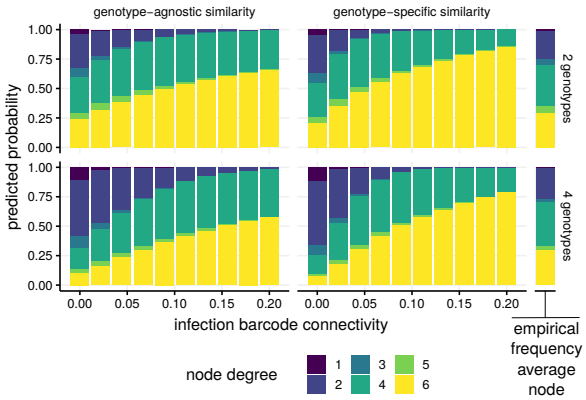

Fig. 4. Node degree predicted based on the infection barcode history in a population with 2 (top) or with 4 (bottom) circulating genotypes. Barcodes similarity matrices are built for infection episodes with the same (right) or any (left) genotypes. Based on the results from multinomial regression models, the host's node degree is predicted from its normalized barcode similarity degree value. The bars on the right show the average node degree frequency from the simulated data on which the statistical model was trained. The higher the node degree in the barcode similarity matrix, the higher its degree in the adjacency matrix.

The spatial correlations between barcode similarity indices and both local and global graph properties also translate to the individual level. From the barcode similarity matrix, we calculate each host's connectivity (i.e. degree) in terms of infection history by summing the host's barcode similarity with respect to all other infected hosts and normalizing appropriately (see Methods). We hypothesise that the infection barcode connectivity within the network of infected hosts can inform the degree of connectivity in the contact network. Multinomial regression tests show that the odds of having higher degree in the contact network (i.e. more neighbors) increases as infection barcode connectivity goes up.

Model predictions show that, with increasing barcode similarity degree, hosts are more likely to have higher network degree (Figure 4), indicating that a host's network degree can be inferred from the infection barcode similarity with all infected hosts in the network.

Discussion

Understanding the properties of host contact networks is key to predicting epidemic spread but raises many practical challenges (13). Some phylodynamics studies have shown that some of these properties can be inferred from virus sequence data (17, 42). Here, we use individual infection histories to gain insights into contact network structure.

A first obstacle was to simulate epidemics on contact networks, while allowing for coinfections, i.e. the simultaneous infection of hosts by multiple parasites (43). Indeed, a direct implementation of the Gillespie algorithm, which is classically used to perform stochastic simulations based on systems of differential equations, is problematic. Indeed, this method being memory-less, it fails to keep track of the infection history. As we show, this generates results that are at odds with the biology. To address this problem, we implemented a non-Markovian version of the Gillespie algorithm following recent developments in statistics (30). As expected, network topology directly impacted multiple infection dynamics, with increased level of clustering leading to increased parasite strain diversity.

Being able to simulate epidemics of multiple infections in an unbiased way, we encountered a second obstacle, which was comparing complex individual infection histories. To address this issue, we captured these histories using barcodes and compared two infection histories using tools from computational topology (37). We could then show that global properties of the network are correlated with these barcodes, i.e. similarity matrices inferred from the barcodes correlate with matrices inferred from the network adjacency matrix. Furthermore, even individual-centered properties such as a host's degree can also be inferred from infection barcodes.

Overall, we provide a proof of principle that multiple infections and infection history can be used to gain insight into host contact network properties. This is biologically sound, given that infections by multiple parasite genotypes are extremely frequent, and realistic. Indeed, for many human infections, there are longitudinal follow-ups with a detailed record of parasite diversity, one of the clearest examples being human papillomaviruses (44).

There are several directions in which our work could be extended. First, we only reported results obtained on random clustered networks. We also obtained similar results on other types of topologies (e.g. Erdős-Rényi, data not shown) but it would be valuable to know whether infection history is more or less informative depending on the type of network considered and if network comparison can be performed. We also assumed that the contact network was static but in reality, it can be highly dynamical. Recent evidence suggests that these dynamics could be detected in sequence data (45) and it would be interesting to explore this with infection histories.

We assumed the life-cycle of the parasite to be that of an SIR with perfect genoytpe-specific immunity. Dynamics with partial cross-immunity or no immunity can be readily implemented. Following many existing studies, e.g. in phylodynamics, we used a neutrality assumption. However, genotypes are known to interact (46) and multiple infections can be a means to detect these interactions (23, 47). In general, exploring richer life-cycles is a promising path for future studies.

In terms of articulating this framework with clinical data, we considered that infection history was based on parasite genotype presence in the host (e.g. PCR detection test) but much more data could be obtained using serologies, i.e. evidence of past presence in the host via antibody detection. An advantage of this approach is that it does not require a detailed longitudinal follow-up. However, the downside is that we then ignore the origin of the infection. Simulation studies could be instrumental in assessing our ability to infer network properties from serological data.

A separate line of future research has to do with the inference *per se*. One possibility could be to use Approximate Bayesian Computation to get a more precise idea of accuracy of the prediction we can make on key network topology parameters. This could also allow us to compare between different classes of networks, e.g. using random forest algorithms (48).

Finally, with the increased power and decreasing cost of sequencing, it may be possible in the future to have both the information about infection history and the sequence data for multiple pathogens. It would then be worth determining whether we can get the best out of the two types of data, infection history being less precise but also less noisy than phylodynamic inference.

Author Affiliations. MIVEGEC (UMR CNRS 5290, IRD 224, UM)
911 Avenue Agropolis
BP 64501
34394 Montpellier cedex 5 - FRANCE

Materials and Methods

Simulation algorithm. We developed an event-driven stochastic model of multiple infections on networks in Python. For the purpose of this simulation study, we considered static, random networks to model contact (i.e. edges with binary weights) between hosts (i.e. nodes). Contact networks were generated using the class of random clustered graphs (49–51) implemented in Python (52). Given a propensity list of edge degrees and triangle degrees as input, this algorithm generates a random graph with predefined average degree and average clustering coefficient. The clustering coefficient of a graph was defined as the local clustering coefficient (i.e. the degree of a node divided by the number of all possible edges in the node's neighbourhood) averaged over all nodes. Degree dispersion was defined as the ratio of degree variance and degree mean. Degree assortativity, i.e. the propensity for nodes in the network that have many connections to be connected to other nodes with many connections was defined following (53).

Upon network initialization, we randomly seeded outbreaks of one infection per genotype, allowing for up to four genotypes per host. Disease dynamics followed Gillespie's stochastic simulation algorithm (SSA) (54, 55) adapted to a non-Markovian setting (56) to incorporate memory-dependent processes. Here in particular, we considered recovery from infection as a process depending on the age of infection. We assumed life-long genotype-specific immunity, i.e. after recovery from an infection with a particular genotype, the host would not be susceptible to this genotype again.Transmission and recovery rates were considered equal for all genotypes and fixed in advance, i.e. genotype frequencies could not vary over time at neutrality without mutation nor migration.

The Gillespie SSA allowed us to simulate disease dynamics by performing regular updates in the values of the rates. For this, at each time point we first created a rate vector $\{r_k\}_{k\in E}$ indexed by the set of all possible events E (recovery from an ongoing infection or acquisition of a new infection by an infectious host in the graph neighborhood). If the node i had spent $t_{i,g}$ time in an infection with genotype g, then the rate of recovery from this infection was given by a Weibull hazard rate function $h(t_{i,g}, \alpha, \mu) = \alpha\mu t_{i,g}^{\alpha-1}$ with shape parameter α and scale parameter μ. If the node i was currently

DRAFT

uninfected with genotype g but had a network neighbour infected with g, then the rate of infection was unity. To better handle the computational workload of multiple infections on networks, we stored the infection histories in an interval tree data structure comprised of the times of onset and end of each infection episode as well as the genotype and the host. We refer to this data structure as infection barcodes (see below). Given the rate vector at present, we first drew an exponentially distributed random variate with rate $\sum_k r_k$ to determine the time increment to the next event. Certain rates r_k depended on the infection age, and it was shown in (56) that drawing exponentially distributed random variates with respect to these rates provides good Markovian approximations of non-Markovian processes, granted the number of events is sufficiently large. We then drew a random variate according to the total probability vector $\{r_k/\sum_k r_k\}_{k\in E}$ to determine the next event. Depending on the event, we updated the list of possible events. In the case of recovery of host i from infection with genotype g, we wrote the end time of the infection episode to the interval tree, removed the recovered infection from the rate vector and added potential infections with genotypes other than g from the host's network neighborhood. In the case of a new infection of the host i with g, we wrote the time of onset of the infection episode to the interval tree, removed all possible edges from g-infected neighbors of i and added possible infection edges from i to all neighbours that had not been or were not currently infected with g. We then updated the rates vector again and proceeded until the epidemic became extinct.

Unless stated otherwise, we simulated disease transmission with four genotypes introduced simultaneously on the giant component of random clustered graphs with 250 nodes at average degree of 4 and average clustering coefficients of 0.36. For memory-less recovery we set $\alpha_g = 1$, and for simulations with memory we chose $\alpha_g = 2$. For a given parameter set, we seeded a random outbreak at the beginning of each simulation with one infection per genotype and ran 50 stochastic replicates until the disease-free state or a pre-defined time horizon was reached. Epidemic outcomes were reported as average and 95% confidence intervals for equidistant time bins.

The source code, configuration files used for simulations and container files will be publicly available on https://gitlab-bioinfo.ird.fr/ETE/MultiNet.

Genotype diversity and multiplicity of infection. We measured genotype diversity during the course of epidemics with $N = 4$ genotypes by Shannon's diversity index (57) defined as $H(t) = \exp\left(\sum_i p_i(t)\log p_i(t)\right)$, where $p_i(t)$ is the frequency of infections with genotype i relative to all infections present at time t. The index $H \in [1, 4]$ is maximised when all genotypes have equal frequency and minimized when only one genotype is present. We emphasize that this index pertains to diversity at the population-level and does not distinguish between a multiple infection within a single host and single infections in several hosts. To highlight the importance of multiple infections, we also considered the multiplicity of infection (MOI) given by the number of genotypes present within infected hosts at a given time. We reported MOI averaged over all infected nodes.

Infection barcodes. We summarized each individual multiple infection history by a barcode, i.e. the set of intervals describing the onset and clearance of infection episodes accumulated within a host during the course of an epidemic. We explicitly included the presence of multiple genotypes in a host's infection history. The notion of barcode first arose in computational topology (58) to succinctly summarize and compare topological properties of metric spaces. Here, we considered the metric space of host barcodes endowed with two complementary notions of distance.

In the first approach, we considered a similarity index (59) such that two hosts with largely overlapping infection episodes tended to be similar to each other, regardless of the respective genotypes. More precisely, a given pair of barcodes between any two hosts $A = \{A_i\}$ and $B = \{B_j\}$ was transformed into a weighted bipartite graph. The nodes of this graph consisted of infection episodes, and the nodes where partitioned according to the two hosts. The edge weight of the graph was given by the overlap length $|A_i \cap B_j|$ between the infection episodes constituting the edge (i, j). To obtain the similarity index $s_1(A, B)$, we calculated the maximum bipartite graph matching of this graph. The similarity index was transformed into a distance by $d_1 = -2s_1 + \sum_i |A_i| + \sum_j |B_j|$.

For the second approach, we considered the overlap length of infection episodes between two hosts of a particular genotype g. By summing over all genotypes we obtained the similarity index $s_2 = \sum_g |A_g \cap B_g|$. This index was transformed into a distance by setting $d_2 = -2s_2 + \sum_i |A_i| + \sum_j |B_j|$.

To determine whether similarity between infection histories implies proximity in the transmission network, we compared the metric space of infection barcodes to the metric space of the network, restricted to nodes that had been infected during the course of the epidemic. For the network, we considered two different distance notions, i.e. the discrete metric based on the binary adjacency matrix and the shortest path distance of the graph.

We used two-sided p-values from the Mantel test (60) between the barcode similarity (resp. distance) matrix and the adjacency or shortest path length matrix respectively to assess for each model the correlation between infection barcode and network topology (61).

To assess whether spatial correlations between similarity matrices also translate to local properties such as a host's node degree (i.e. number of network neighbors), we performed a multinomial regression analysis. For a given simulation with maximum length L, N_g different genotypes and $D + 1$ nodes in the giant component, a host n's normalized infection barcode connectivity is defined by

$$\hat{s}_i(n) = (LN_gD)^{-1} \sum_m s_i(n, m) \in [0, 1]$$

for $i = 1, 2$, i.e. the sum of all barcode similarity values within the network of infected hosts relative to the maximum possible barcode similarity. The multinomial logistic regression model was defined as

$$\log \frac{\mathbb{P}(\text{degree} = k)}{\mathbb{P}(\text{degree} = 1)} = \beta_0 + \beta_1 \hat{s}_i + \beta_2 N_g$$

where β_0 is the intercept, β_1 is the coefficient for continuous variable of the normalized barcode similarity degree, and β_2 is the coefficient for the categorical variable of $N_g = 2$ relative to the base level $N_g = 4$. We evaluated the model for genotype-agnostic $\hat{s}_1$ and genotype-specific $\hat{s}_2$ similarity degrees. Significance for each coefficient was assessed by p-values from two-sided Z-tests. We used the Akaike Information Criterion (AIC) to compare with ordered logistic models

$$\log \frac{\mathbb{P}(\text{degree} = k)}{\mathbb{P}(\text{degree} < k)} = \beta_0 + \beta_1 \hat{s}_i + \beta_2 N_g$$

which yielded similar predicted values at consistently higher AIC (data not shown).

ACKNOWLEDGMENTS. This work has been funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (EVOLPROOF, grant agreement No 648963 to SA). The authors acknowledge additional support from the CNRS, the IRD and the itrop HPC (South Green Platform in Montpellier for providing HPC resources that have contributed to the research results reported within this study (https://bioinfo.ird.fr/).

1. R. M. Anderson and R. M. May. *Infectious Diseases of Humans. Dynamics and Control.* Oxford University Press, Oxford, 1991.
2. O. Diekmann and JA Heesterbeek. *Mathematical epidemiology of infectious diseases: model building, analysis, and interpretation.* Wiley, New York, 2000.
3. M. J. Keeling and P. Rohani. *Modeling infectious diseases in humans and animals.* Princeton University Press, 2008.
4. R. M. May and R. M. Anderson. Transmission dynamics of hiv infection. *Nature*, 326(6109): 137–42, 1987.
5. H K Alexander and T Day. Risk factors for the evolutionary emergence of pathogens. *J R Soc Interface*, 7(51):1455–74, 2010. .
6. Gabriel E. Leventhal, Alison L. Hill, Martin A. Nowak, and Sebastian Bonhoeffer. Evolution and emergence of infectious diseases in theoretical and real-world networks. *Nat Commun*, 6:6101, 2015. ISSN 2041-1723. .
7. M. J. Keeling. The effects of local spatial structure on epidemiological invasions. *Proc. R. Soc. Lond. B*, 266:859–867, 1999.
8. M. E. J. Newman. Spread of epidemic disease on networks. *Phys. Rev. E*, 66(1):16128, 2002. .
9. Matt J. Keeling and Ken T. D. Eames. Networks and epidemic models. *J R Soc Interface*, 2 (4):295–307, 2005. ISSN 1742-5689. .

DRAFT

10. Gregory M. Ames, Dylan B. George, Christian P. Hampson, Andrew R. Kanarek, Cayla D. McBee, Dale R. Lockwood, Jeffrey D. Achter, and Colleen T. Webb. Using network properties to predict disease dynamics on human contact networks. *Proc R Soc B*, 278(1724):3544–3550, 2011. ISSN 0962-8452. .
11. Lorenzo Pellis, Frank Ball, Shweta Bansal, Ken Eames, Thomas House, Valerie Isham, and Pieter Trapman. Eight challenges for network epidemic models. *Epidemics*, 10:58–62, 2015. ISSN 1755-4365. .
12. Laurent Hébert-Dufresne and Benjamin M. Althouse. Complex dynamics of synergistic coinfections on realistically clustered networks. *Proc Natl Acad Sci USA*, 112(33):10551–10556, 2015. ISSN 0027-8424. .
13. K. Eames, S. Bansal, S. Frost, and S. Riley. Six challenges in measuring contact networks for use in modelling. *Epidemics*, 10:72–77, 2015. ISSN 1755-4365. .
14. Istvan Z Kiss, Darren M Green, and Rowland R Kao. Disease contact tracing in random and clustered networks. *Proceedings of the Royal Society B: Biological Sciences*, 272(1570): 1407–1414, July 2005. . URL https://doi.org/10.1098/rspb.2005.3092.
15. Vittoria Colizza, Alain Barrat, Marc Barthélemy, and Alessandro Vespignani. The role of the airline transportation network in the prediction and predictability of global epidemics. *Proc Natl Acad Sci USA*, 103(7):2015–20, 2006. .
16. Nathan Eagle, Alex (Sandy) Pentland, and David Lazer. Inferring friendship network structure by using mobile phone data. *Proc Natl Acad Sci USA*, 106(36):15274–15278, 2009. ISSN 0027-8424. .
17. Gabriel E Leventhal, Roger Kouyos, Tanja Stadler, Viktor von Wyl, Sabine Yerly, Jürg Böni, Cristina Cellerai, Thomas Klimkait, Huldrych F Günthard, and Sebastian Bonhoeffer. Inferring epidemic contact structure from phylogenetic trees. *PLoS Comput Biol*, 8(3):e1002413, 2012. .
18. Eugenio Valdano, Chiara Poletto, Armando Giovannini, Diana Palma, Lara Savini, and Vittoria Colizza. Predicting Epidemic Risk from Past Temporal Contact Data. *PLoS Comput Biol*, 11 (3):e1004152, 2015. ISSN 1553-7358. .
19. Jos WR Twisk. *Applied longitudinal data analysis for epidemiology: a practical guide*. Cambridge University Press, 2013.
20. Jonathan J Juliano, Kimberly Porter, Victor Mwapasa, Rithy Sem, William O Rogers, Frédéric Ariey, Chansuda Wongsrichanalai, Andrew Read, and Steven R Meshnick. Exposing malaria in-host diversity and estimating population diversity by capture-recapture using massively parallel pyrosequencing. *Proc Natl Acad Sci USA*, 107(46):20138–43, 2010. .
21. Anil K. Chaturvedi, Hormuzd A. Katki, Allan Hildesheim, Ana Cecilia Rodríguez, Wim Quint, Mark Schiffman, Leen-Jan Van Doorn, 2 Sholom Wacholder Carolina Porras, Paula Gonzalez, Mark E. Sherman, Rolando Herrero, and the CVT Group. Human papillomavirus infection with multiple types: Pattern of coinfection and risk of cervical disease. *J Infect Dis*, 203:910–920, 2011. .
22. D. J. Shaw and A. P. Dobson. Patterns of macroparasite abundance and aggregation in wildlife populations: a quantitative review. *Parasitology*, 111(S1):S111–S133, 1995. ISSN 1469-8161. .
23. Samuel Alizon, Carmen Lia Murall, Emma Saulnier, and Mircea T Sofonea. Detecting within-host interactions from genotype combination prevalence data. *bioRxiv*, 2018. . URL https://www.biorxiv.org/content/early/2018/10/05/256586.
24. C. OF. Buckee, K. Koelle, M. J. Mustard, and S. Gupta. The effects of host contact network structure on pathogen diversity and strain structure. *Proceedings of the National Academy of Sciences*, 101(29):10839–10844, jul 2004. . URL https://doi.org/10.1073/pnas.0402000101.
25. J. C. Miller. Cocirculation of infectious diseases on networks. *Phys Rev E Stat Nonlin Soft Matter Phys*, 87(6):060801, Jun 2013.
26. Chiara Poletto, Sandro Meloni, Vittoria Colizza, Yamir Moreno, and Alessandro Vespignani. Host mobility drives pathogen competition in spatially structured populations. *PLoS Computational Biology*, 9(8):e1003169, August 2013. . URL https://doi.org/10.1371/journal.pcbi.1003169.
27. Chiara Poletto, Sandro Meloni, Ashleigh Van Metre, Vittoria Colizza, Yamir Moreno, and Alessandro Vespignani. Characterising two-pathogen competition in spatially structured environments. *Scientific Reports*, 5(1), January 2015. . URL https://doi.org/10.1038/srep07895.
28. Faryad Darabi Sahneh and Caterina Scoglio. Competitive epidemic spreading over arbitrary multilayer networks. *Physical Review E*, 89(6), June 2014. . URL https://doi.org/10.1103/physreve.89.062817.
29. Francesco Pinotti, Éric Fleury, Didier Guillemot, Pierre-Yves Böelle, and Chiara Poletto. Host contact dynamics shapes richness and dominance of pathogen strains. September 2018. . URL https://doi.org/10.1101/428185.
30. Marian Boguña, Luis F. Lafuerza, R Toral, and M A Serrano. Simulating non-Markovian stochastic processes. *Phys Rev E*, 90(4):042108, 2014. ISSN 2470-0053. .
31. NORMAN T. J. BAILEY. ON ESTIMATING THE LATENT AND INFECTIOUS PERIODS OF MEASLES. *Biometrika*, 43(1-2):15–22, 1956. . URL https://doi.org/10.1093/biomet/43.1-2.15.
32. M. Eichner and K. Dietz. Transmission potential of smallpox: Estimates based on detailed data from an outbreak. *American Journal of Epidemiology*, 158(2):110–117, July 2003. . URL https://doi.org/10.1093/aje/kwg103.
33. W. Sama, K. Dietz, and T. Smith. Distribution of survival times of deliberate plasmodium falciparum infections in tertiary syphilis patients. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 100(9):811–816, September 2006. . URL https://doi.org/10.1016/j.trstmh.2005.11.001.
34. Hormuzd A. Katki, Li C. Cheung, Barbara Fetterman, Philip E. Castle, and Rajeshwari Sundaram. A joint model of persistent human papilloma virus infection and cervical cancer risk: implications for cervical cancer screening. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 178(4):903–923, March 2015. . URL https://doi.org/10.1111/rssa.12101.
35. Michael T. Bretscher, Nicolas Maire, Nakul Chitnis, Ingrid Felger, Seth Owusu-Agyei, and Tom Smith. The distribution of plasmodium falciparum infection durations. *Epidemics*, 3(2): 109–118, jun 2011. . URL https://doi.org/10.1016/j.epidem.2011.03.002.
36. Benjamin J Cowling, Lianmei Jin, Eric HY Lau, Qiaohong Liao, Peng Wu, Hui Jiang, Tim K Tsang, Jiandong Zheng, Vicky J Fang, Zhaorui Chang, Michael Y Ni, Qian Zhang, Dennis KM Ip, Jianxing Yu, Yu Li, Liping Wang, Wenxiao Tu, Ling Meng, Joseph T Wu, Huiming Luo, Qun Li, Yuelong Shu, Zhongjie Li, Zijian Feng, Weizhong Yang, Yu Wang, Gabriel M Leung, and Hongjie Yu. Comparative epidemiology of human infections with avian influenza a h7n9 and h5n1 viruses in china: a population-based study of laboratory-confirmed cases. *The Lancet*, 382(9887):129–137, jul 2013. . URL https://doi.org/10.1016/s0140-6736(13)61171-x.
37. Herbert Edelsbrunner and John Harer. *Computational Topology - an Introduction*. American Mathematical Society, 2010. ISBN 978-0-8218-4925-5.
38. M. J. Keeling. Disease extinction and community size: Modeling the persistence of measles. *Science*, 275(5296):65–67, January 1997. . URL https://doi.org/10.1126/science.275.5296.65.
39. Alun L. Lloyd. Realistic distributions of infectious periods in epidemic models: Changing patterns of persistence and dynamics. *Theoretical Population Biology*, 60(1):59–71, August 2001. . URL https://doi.org/10.1006/tpbi.2001.1525.
40. K. J. GOUGH. The estimation of latent and infectious periods. *Biometrika*, 64(3):559–565, 1977. . URL https://doi.org/10.1093/biomet/64.3.559.
41. Scott Greenhalgh and Troy Day. Time-varying and state-dependent recovery rates in epidemiological models. *Infectious Disease Modelling*, 2(4):419–430, November 2017. . URL https://doi.org/10.1016/j.idm.2017.09.002.
42. David A. Rasmussen, Roger Kouyos, Huldrych F. Günthard, and Tanja Stadler. Phylodynamics on local sexual contact networks. *PLOS Computational Biology*, 13(3):e1005448, March 2017. . URL https://doi.org/10.1371/journal.pcbi.1005448.
43. Mircea T. Sofonea, Samuel Alizon, and Yannis Michalakis. Exposing the diversity of multiple infection patterns. *Journal of Theoretical Biology*, 419:278–289, April 2017. . URL https://doi.org/10.1016/j.jtbi.2017.02.011.
44. R. M. Brotman, M. D. Shardell, P. Gajer, J. K. Tracy, J. M. Zenilman, J. Ravel, and P. E. Gravitt. Interplay between the temporal dynamics of the vaginal microbiota and human papillomavirus detection. *Journal of Infectious Diseases*, 210(11):1723–1733, June 2014. . URL https://doi.org/10.1093/infdis/jiu330.
45. Cornelia Metzig, Oliver Ratmann, Daniela Bezemer, and Caroline Colijn. Phylogenies from dynamic networks. *PLOS Computational Biology*, 15(2):e1006761, February 2019. . URL https://doi.org/10.1371/journal.pcbi.1006761.
46. Nicole Mideo. Parasite adaptations to within-host competition. *Trends in Parasitology*, 25(6): 261–268, June 2009. . URL https://doi.org/10.1016/j.pt.2009.03.001.
47. Irene Man, Kari Auranen, Jacco Wallinga, and Johannes A. Bogaards. Capturing multiple-type interactions into practical predictors of type replacement following human papillomavirus vaccination. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 374 (1773):20180298, May 2019. . URL https://doi.org/10.1098/rstb.2018.0298.
48. Pierre Pudlo, Jean-Michel Marin, Arnaud Estoup, Jean-Marie Cornuet, Mathieu Gautier, and Christian P. Robert. Reliable ABC model choice via random forests. *Bioinformatics*, 32(6): 859–866, November 2015. . URL https://doi.org/10.1093/bioinformatics/btv684.
49. Shweta Bansal, Shashank Khandelwal, and Lauren Ancel Meyers. Exploring biological network structure with clustered random networks. *BMC Bioinformatics*, 10(1), dec 2009. . URL https://doi.org/10.1186/1471-2105-10-405.
50. Joel C. Miller. Percolation and epidemics in random clustered networks. *Physical Review E*, 80(2), aug 2009. . URL https://doi.org/10.1103/physreve.80.020901.
51. M. E. J. Newman. Random graphs with clustering. *Physical Review Letters*, 103(5), jul 2009. . URL https://doi.org/10.1103/physrevlett.103.058701.
52. Aric A. Hagberg, Daniel A. Schult, and Pieter J. Swart. Exploring network structure, dynamics, and function using networkx. In Gaël Varoquaux, Travis Vaught, and Jarrod Millman, editors, *Proceedings of the 7th Python in Science Conference*, pages 11 – 15, Pasadena, CA USA, 2008.
53. M. E. J. Newman. Assortative mixing in networks. *Physical Review Letters*, 89(20), October 2002. . URL https://doi.org/10.1103/physrevlett.89.208701.
54. Daniel T Gillespie. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *Journal of Computational Physics*, 22(4):403–434, dec 1976. . URL https://doi.org/10.1016/0021-9991(76)90041-3.
55. Daniel T. Gillespie. Exact stochastic simulation of coupled chemical reactions. *The Journal of Physical Chemistry*, 81(25):2340–2361, dec 1977. . URL https://doi.org/10.1021/j100540a008.
56. M. Boguña, L. F. Lafuerza, R. Toral, and M. A. Serrano. Simulating non-Markovian stochastic processes. *Phys Rev E Stat Nonlin Soft Matter Phys*, 90(4):042108, Oct 2014.
57. E. Kathryn Morris, Tancredi Caruso, François Buscot, Markus Fischer, Christine Hancock, Tanja S. Maier, Torsten Meiners, Caroline Müller, Elisabeth Obermaier, Daniel Prati, Stephanie A. Socher, Ilja Sonnemann, Nicole Wäschke, Tesfaye Wubet, Susanne Wurst, and Matthias C. Rillig. Choosing and using diversity indices: insights for ecological applications from the german biodiversity exploratories. *Ecology and Evolution*, 4(18):3514–3524, August 2014. . URL https://doi.org/10.1002/ece3.1155.
58. F. Chazal and B. Michel. An introduction to Topological Data Analysis: fundamental and practical aspects for data scientists. *ArXiv e-prints*, 2017.
59. Gunnar Carlsson, Afra Zomorodian, Anne Collins, and Leonidas Guibas. Persistence barcodes for shapes. In *Proceedings of the 2004 Eurographics/ACM SIGGRAPH Symposium on Geometry Processing*, SGP '04, pages 124–135, New York, NY, USA, 2004. ACM. ISBN 3-905673-13-4. . URL http://doi.acm.org/10.1145/1057432.1057449.
60. J. A. Diniz-Filho, T. N. Soares, J. S. Lima, R. Dobrovolski, V. L. Landeiro, M. P. de Campos Telles, T. F. Rangel, and L. M. Bini. Mantel test in population genetics. *Genet. Mol. Biol.*, 36(4):475–485, 2013.
61. Sarah C. Goslee and Dean L. Urban. The ecodist package for dissimilarity-based analysis of ecological data. *Journal of Statistical Software*, 22:1–19, 2007.

DRAFT