Quantifying transmission dynamics of acute hepatitis C virus infections in a heterogeneous population using sequence data

Gonché Danesh[1],~~Victor Virlogeux~~Victor Virlogeux[2],Christophe Ramière[3]~~¤~~,Caroline Charre[3], Laurent Cotte[4‡], Samuel Alizon[1‡]

1 MIVEGEC (UMR CNRS 5290, IRD, UM), Montpellier, France

2 Clinical Research Center, Croix-Rousse Hospital, Hospices Civils de Lyon, France

3 Virology Laboratory, Croix-Rousse Hospital, Hospices Civils de Lyon, France

4 Infectious Diseases Department, Croix-Rousse Hospital, Hospices Civils de Lyon, France

~~These authors contributed equally to this work.~~

‡These authors ~~also~~ contributed equally to this work.

* Corresponding author: gonche.danesh@ird.fr

Abstract

Opioid substitution and syringes exchange programs have drastically reduced hepatitis C virus (HCV) spread in France but HCV sexual transmission in men having sex with men (MSM) has recently ~~arose~~ arisen as a significant public health concern. The fact that the virus is transmitting in a heterogeneous population, with 'new' and 'classical' hosts, makes prevalence and incidence rates poorly informative. However, additional insights can be gained by analyzing ~~dated~~ virus phylogenies inferred from dated ~~virus~~ genetic sequence data. Here, using ~~such~~ a phylodynamics approach based on Approximate Bayesian Computation, we estimate key epidemiological parameters of an ongoing HCV epidemic in MSM in Lyon (France). We show that this ~~epidemics in MSM~~ new epidemics is largely independent from the 'classical' HCV epidemics and that its doubling time is one order of magnitude lower (55.6 days *versus* 511 days). These results have practical implications for HCV control and ~~open new perspective for using~~ illustrate the additional information provided by virus genomics in public health.

Background

~~The burden of~~ It is estimated that 71 million people worldwide suffer from chronic hepatitis C virus (HCV) ~~infection is currently estimated to 71 million infections worldwide~~ infections [?, ?]. The ~~virus being exclusively a human agent, the~~ World Health Organisation (WHO) and several countries have issued recommendations towards ~~its~~ the 'elimination' ~~. This means the absence of a significant transmission in a given epidemiological context and is defined by a~~ of this virus, which they define as an 80% reduction in new chronic infections and a 65% decline in liver mortality by 2030 [?]. HIV-HCV coinfected patients are ~~considered a key population~~ targeted with priority because of the shared ~~routes of transmission~~ transmission routes between the two viruses [?] and because of the increased ~~severity of chronic HCV infection~~ virulence of HCV in coinfections [?, ?, ?]. ~~This population was therefore targeted with priority, leading to high treatment uptake in countries with affordable access to treatments. This was complemented by successful harm reductions interventionssuch as needle-syringes exchange and opiates~~ Successful harm reduction interventions, such as needle-syringe exchange and opiate substitution programs, as well as ~~the~~ a high level of enrolment into care of HIV-infected patients~~. As a result,~~ , have led to a drastic drop in the prevalence of active HCV infections ~~drastically dropped during the recent years~~ in HIV-HCV coinfected patients in several European countries ~~, increasing the hope that HCV elimination was an attainable goal [?, ?, ?, ?].~~

~~Unfortunately, sexual transmission of HCV in HIV-infected~~ during the recent years [?, ?, ?, ?] . Unfortunately, this elimination goal is challenged by the emergence of HCV sexual transmission, especially among men having sex with men (MSM)~~recently arose as a significant phenomenon,~~ . This trend is reported to be driven by unprotected sex~~and~~ , drug use in the context of sex ('chemsex')~~and by~~ , and potentially traumatic practices such as fisting ~~and sharing sextoys [?, ?, ?] . This is the case in~~ [?, ?, ?]. In area of Lyon (France), ~~where~~ HCV incidence has been shown to increase concomitantly with a shift in the profile of infected hosts [?]. Understanding and quantifying this recent increase is the main goal of this study.

Several modeling studies have highlighted the ~~difficulties to control HCV infection~~ difficulty to control the spread of HCV infections in HIV-infected MSM in the absence of harm reduction interventions [?, ?]. Furthermore, we recently described the spread of HCV from HIV-infected to HIV-negative MSM, using ~~or not~~ HIV pre-exposure prophylaxis (PrEP) ~~, through sharing of~~ or not, through shared high-risk practices ~~between these populations [?]. This resulted in~~ [?] . More generally, an alarming incidence of acute HCV ~~infection~~ infections in both HIV-infected and PrEP-using MSM was reported in France in 2016-2017 [?]. Additionally, while PrEP-using MSM are regularly screened for HCV, those who are HIV-negative and do not use PrEP may remain undiagnosed and untreated for years. ~~Since~~ In general, we know little about the population size and ~~the~~ practices of HIV-negative MSM who do not use PrEP~~, these recent~~ . All these epidemiological events could jeopardize the goal of HCV elimination by creating a large pool of infected and undiagnosed patients, ~~pursuing high-risk practices that~~ which could fuel new infections in intersecting populations. Furthermore, the epidemiological dynamics of HCV infection have mostly been studied in intravenous drug users (IDU) [?, ?, ?, ?] and in the general population [?, ?].

Results from these populations are not easily transferable to other populations, which calls for a better understanding of the epidemiological characteristics of HCV sexual transmission in MSM.

Given the lack of knowledge about the focal population driving the increase in HCV incidence, we ~~investigated~~ analyse virus sequence data ~~using~~ with phylodynamics methods. This research field has been blooming over the last decade ~~[?, ?, ?]~~ and hypothesizes that the way rapidly evolving viruses spread leaves 'footprints' in their genomes [?, ?, ?]. By combining ~~epidemiological modeling~~mathematical modelling, statistical analyses and phylogenies of infections, where each leaf corresponds to the virus sequence isolated from a patient, current methods can ~~estimate key transmission~~ infer key parameters of viral epidemics. This framework has been successfully applied to other HCV epidemics [?, ?, ?, ?]~~but the epidemics we study is particularly~~, but the ongoing one in Lyon is challenging to analyze because the focal population is heterogeneous, with 'classical' hosts (typically ~~HIV-infected patients~~ HIV-negative patients infected through nosocomial transmission or with a history of opioid intravenous drug use or blood transfusion) and 'new' hosts (both HIV-infected and HIV-negative MSM, detected during or shortly after acute HCV infection phase, potentially using recreational drugs such as cocaine or cathinones). ~~To address this issue, we used a framework based on~~ Our phylodynamics analysis relies on an Approximate Bayesian Computation (ABC~~)~~, [?] ~~)~~ framework that was recently developed and validated [?]. ~~We implemented~~

Assuming an epidemiological model with two host types~~(~~, 'classical' and 'new' ~~), where each infection can generate secondary infections before ending~~ (see the Methods)~~. By analyzing~~, we use dated virus sequences ~~, we estimated~~ to estimate the date of onset of the HCV epidemics in ~~the~~ 'classical' ~~hosts, and in the~~ and 'new' hosts, the level of mixing between ~~the~~ hosts types, and, for each host type, the duration of the infectious period and the ~~basic~~ effective reproduction ratio (i.e. the number of secondary infections, [?]). ~~This allowed us to show~~ We find that the doubling time of the epidemics is one order of magnitude lower in 'new' ~~hosts is dramatically higher than that~~ than in 'classical' hosts, therefore emphasising the urgent need for public health action.

Results

The ~~time-scaled~~ phylogeny inferred from the dated virus sequences ~~reveals~~ shows that 'new' hosts (in red) tend to ~~cluster together~~ be grouped in clades (Figure 1). This pattern suggests a high ~~level of assortativity~~ degree of assortativity in the epidemics (i.e.~~each infected host~~ hosts tends to infect hosts from the same type). ~~Furthermore, the estimate for the root of the phylogeny, that is the onset of the epidemics in the studied area, is in the early 1980s, which appears consistent with epidemiological data~~The ABC phylodynamics approach allows us to go beyond a visual description and to quantify several epidemiological parameters.

As for any ~~bayesian~~ Bayesian inference method, we need to assume a prior distribution for each parameter~~(~~. These priors, shown in grey in Figure 2~~) in order to infer posterior distributions (in red in Figure 2). Priors were voluntarily assumed~~, are voluntarily designed to be large and uniformly distributed so as to be as little informative as possible. ~~The only exception was~~ One exception is the date of onset of the epidemics, for which we ~~used~~ use as a prior the output of the phylogenetic analysis ~~as a prior. For~~ conducted in Beast (see the Methods). We also assume the date of the

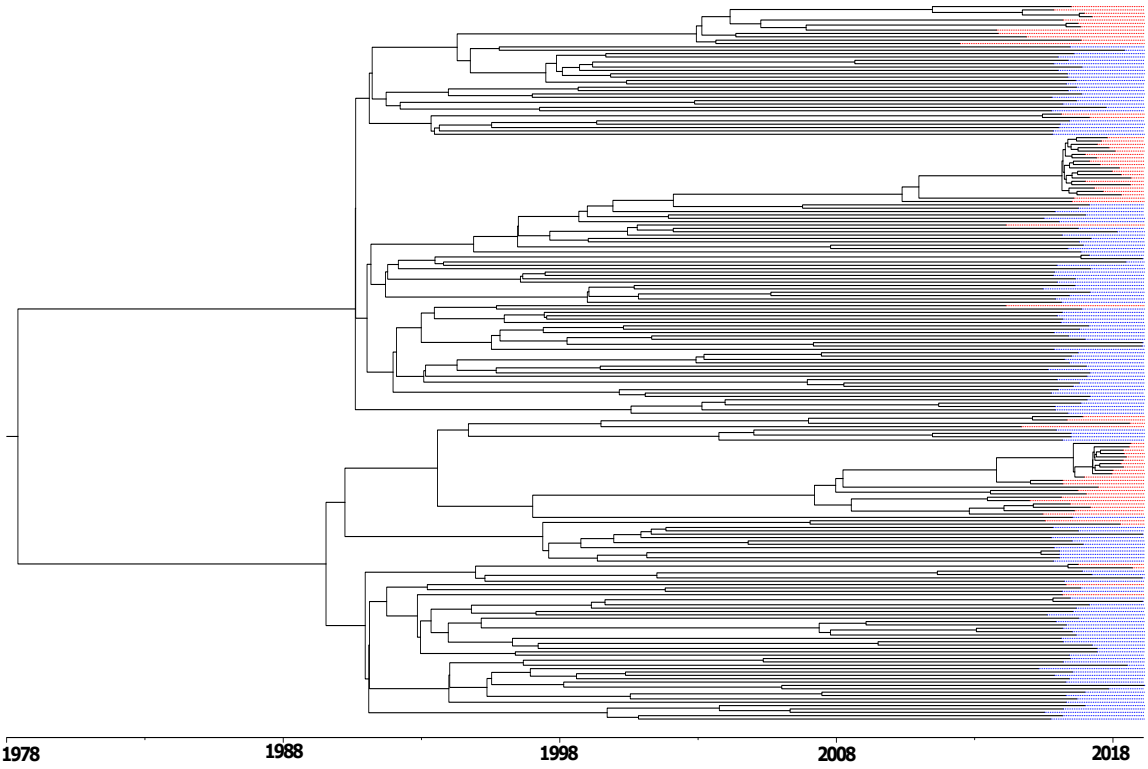


Fig 1. Phylogeny of HCV infections in the area of Lyon (France). 'Classical' hosts are in blue and 'new' hosts are in red. Sampling events correspond to the end of black branches. The phylogeny was estimated using maximum-likelihood methods (PhyML) and then rooted in time using ~~bayesian~~ Bayesian inference (Beast2). See the Methods for additional details.

~~second epidemics , we assumed that it took place after~~ 'new' hosts epidemics to be posterior to 1997 based on epidemiological data. ~~The width of the posterior distribution indicates our ability to infer a parameter.~~

The ~~ABC phylodynamics approach allows us to go beyond a visual description and to quantify several epidemiological parameters. For instance, we can narrow down the estimation~~ inference method converges towards posterior distributions for each parameter, which are shown in red in Figure 2. The estimate for the origin of the epidemic ~~(~~in 'classical' hosts ~~only) to 1976[1969; 1980]~~ is $t_0 = 1977$ [1966; 1981] (numbers in brackets indicate the 95% Highest Posterior Density, or HPD). ~~The epidemic in the second host typeis estimated~~ For the 'new' host type, we estimate the epidemic to have started in ~~2001[1998; 2005]~~$t_2 = 2003$ [2000; 2005].

~~Regarding~~ We find the level of assortativity between host types ~~, that is the extent to which a host of a given type interacts with hosts of the same type, we estimate a_1 to be 0.96[0.86; 0.99] and a_2 to be 0.86[0.72; 0.99]~~to be high for 'classical' ($a_1 = 0.97$ [0.91; 0.99]) as well as for 'new' hosts ($a_2 = 0.88$ [0.70; 0.99]). Therefore, hosts ~~appear to preferentially interact with~~ mainly infect hosts from the same type and this effect seems even more pronounced for 'classical' hosts.

The phylodynamics approach also allows us to infer the duration of the infectious period ~~. Here,~~

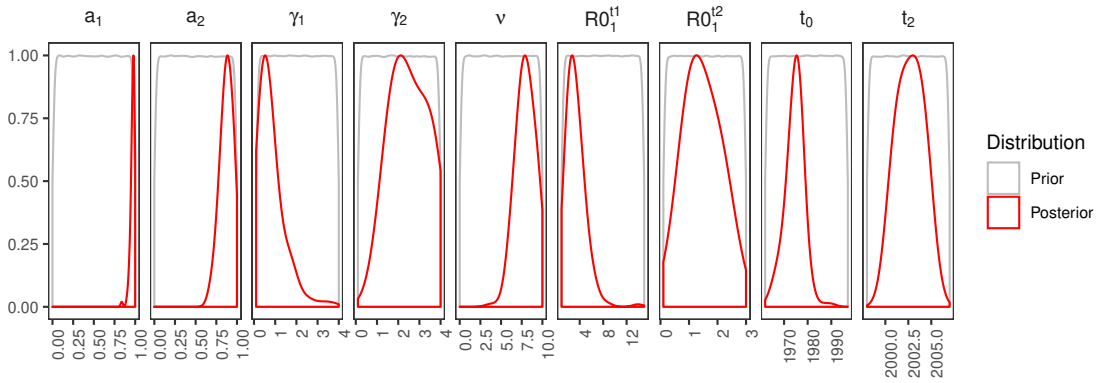


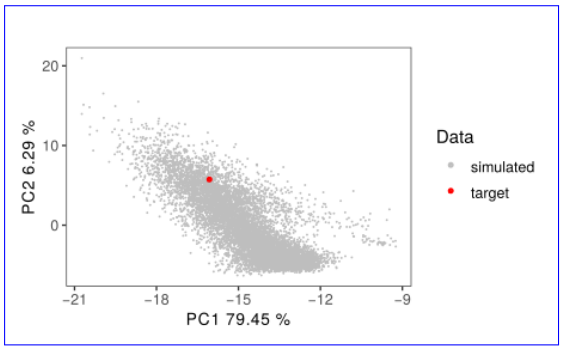
Fig 2. Parameter prior and posterior distributions. Prior distributions are in grey and posterior distributions ~~of the regression-ABC~~ inferred by ABC are in red. The thinner the posterior distribution, the more accurate the inference.

~~assuming~~ for each host type. Assuming that this parameter ~~remains constant for a given host type~~ does not vary over time, we estimate it to be ~~1.7 years [0.46; 9.17]~~ 1.2 years [0.40; 7.69] for 'classical' hosts (parameter $1/\gamma_1$) and 0.4 years ~~[0.26; 0.65]~~ [0.25; 0.78] for 'new' hosts (parameter $1/\gamma_2$).

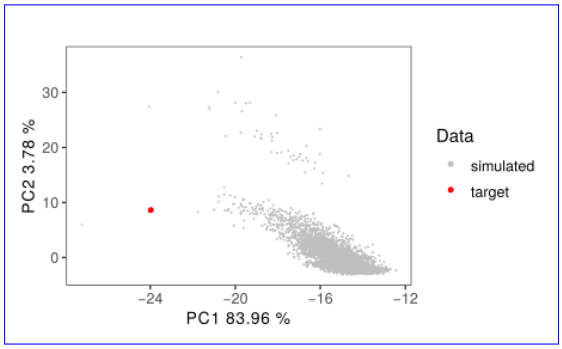
~~The basic~~ Regarding effective reproduction numbers, i.e. the number of secondary infections caused by a given host over ~~the~~ its infectious period, ~~was estimated~~ we estimate that of 'classical' hosts to have decreased from ~~5.94[3.24; 8.61] to 1.80[1.12; 2.48] for 'classical' hosts,~~ $R_0^{(1),t_1} = 3.29$ [1.2; 6.63] to $R_0^{(1),t_2} = 1.47$ [0.37; 2.67] after the introduction of the third generation HCV test in 1997. ~~We also estimate that~~ The inference on the differential transmission parameter indicates that HCV transmission rate is $\nu = 7.97$ [6.01; 9.90] times greater from 'new' hosts ~~transmit HCV 6.50[2.56; 9.81] times more than~~ than from 'classical' hosts~~(parameter ν). Using all these inferences, we can calculate the~~. By combining these results (see the Methods), we estimate the effective reproduction number in 'new' hosts ~~, $R_0^{(2),t_3}$ (see the Methods), which is 2.35[0.55; 8.05]~~ to be $R_0^{(2),t_3} = 2.9$ [0.81; 6.26].

To better ~~show~~ apprehend the differences between the two host types, we compute the ~~epidemics doubling times~~ epidemic doubling time (t_D), which is the time ~~it takes~~ for an infected population to double in size~~,~~. t_D is computed for each type of host, assuming ~~a full assortativity~~ complete assortativity (see the Methods). We find that ~~since 1997, the $t_D^{(1),t2}$ could be estimated to 511.0 days ([0.58; 10.13] years)~~ for the 'classical' hosts, ~~whereas the $t_D^{(2),t3}$ was estimated to 55.56 days ([0; 3.51] years) for the 'new' hosts. Before~~ before 1997 $t_D^{(1),t1} \approx 8$ months ([0.1; 2.63] years). After 1997, the ~~$t_D^{(1),t1}$ was estimated to 83~~ pace decreases with a doubling time of $t_D^{(1),t2} \approx 1.75$ years ([0; 28.55] years). For the epidemics in the 'new' hosts, we estimate that $t_D^{(2),t3} \approx 51$ days (~~[0.05; 1.60] years) for the 'classical' hosts. We show the densities of~~ [0; 2.73] years). Distributions for theses three doubling times are shown in Supplementary Figure S2.

~~In~~ Supplementary Figure S3 ~~, we show~~ shows the correlations between parameters ~~in~~ based on the posterior distributions~~(Figure 3)~~. We mainly find that the R_0 in 'classical' hosts after the introduction of the third generation of HCV detection tests (i.e. $R_0^{(1),t_2}$) is negatively correlated to ν and positively correlated to γ_2. ~~This makes sense because a rapid growth of the epidemic~~



(a)



(b)

Fig 3. Parameteric bootstrap illustration. Principal Component Analysis (PCA~~graph~~) graphs where each dot represents a vector of summary statistics of a ~~data~~dataset. The ~~1,000~~ 5,000 simulated data are in grey, and the target data is in red. Panel (a) shows the PCA graph using the HPD distribution. Panel (b) shows the PCA graph using a uniform distribution drawn from the 95% HPD distribution.

In other words, if the the epidemic spreads rapidly in 'classical' hosts~~imposes a lower growth~~, it requires a slower spread in 'new' hosts to explain the phylogeny. $R_0^{(1),t_2}$ is also slightly negatively correlated to γ_1, which probably comes from the fact that ~~epidemics with the same~~ for a given R_0 ~~but~~, epidemics with a longer infection duration have a lower doubling time and therefore a weaker epidemiological impact. Overall, these correlations do not affect our main results, especially the pronounced difference in infection periods (γ_1 and γ_2).

To validate these results, we ~~performed~~ perform a parametric bootstrap analysis by simulating phylogenies using ~~our~~ the resulting posterior distributions to determine whether these are similar to the target dataset (see the Methods). In Figure ~~??~~3(a), we see that the target data in red, i.e. the summary statistics from the phylogeny shown in Figure 1, lies in the middle of the phylogenies simulated using the posterior data. ~~Even if~~ If we use the 95% HPD of the posterior but assume a uniform distribution instead of the true posterior distribution, we find that the target phylogeny lies outside the cloud of simulations (see ~~Supplementary figure S4~~Figure 3(b)). These results confirm that the posterior distributions we infer are highly informative~~regarding the phylogeny shape~~.

~~Finally, to further validate the accuracy~~ To further explore the robustness of our inference

method, we ~~used~~ use simulated data to perform a 'leave one out' cross-validation (see the Methods). As shown in Supplementary Figure S5, the relative error made for each parameter inference is limited and comparable to what ~~was~~ is found using a simpler model [**?**]. Two exceptions are the rate at which 'new' hosts clear the infection (γ_2) and their level of assortativity (a_2). ~~However, this is likely to be due to the constraint imposed by the shape of the target phylogeny itself rather than by the method. In general, this cross-validation goes beyond the scope of this epidemiological model because, for instance, assortativity values can vary between 0 and 1, whereas for the phylogenetic structure we studied with a high degree of clustering, we expect it should be close to 1~~This is likely a consequence of our choice of summary statistics, which is optimised to analyse a phylogeny with a high degree of assortativity (high values of a_1 and a_2).

Finally, to evaluate the impact of phylogenetic reconstruction uncertainty, we perform a supplementary analysis using 10 additional trees from the Beast posterior distribution. In Supplementary figure S6, we show that the posterior distributions estimated by our ABC method are qualitatively similar with all these trees.

Discussion

Over the last years, ~~an increase in HCV incidence has been witnessed in~~ the area of Lyon (France) ~~, which involves both~~ witnessed an increase in HCV incidence both in HIV-positive and HIV-negative populations of men having sex with men (MSM)~~and~~ [**?**]. This increase appears to be driven by sexual transmission ~~. Similar trends have been described~~ and echoes similar trends in Amsterdam [**?**] and ~~, more recently,~~ in Switzerland [**?**]. ~~Achieving a~~ A quantitative analysis of ~~this epidemics is required~~ the epidemic is necessary to optimise public health interventions~~but extremely~~ . Unfortunately, this is challenging because the monitoring of the population at risk is limited and because classical tools in quantitative epidemiology~~such as~~ , especially incidence time series, are poorly informative ~~in~~ with such a heterogeneous population. To ~~address this issue, we analysed virus sequence data~~circumvent this problem, we used HCV sequence data, which we analysed using phylodynamics. In order to account for host heterogeneity, we extended and validated an existing ~~framework relying on~~ Approximate Bayesian Computation framework [**?**].

From a public health point of view, ~~these~~ our results have two major implications. First, ~~there is a strong assortativity within the~~ we find a strong degree of assortativity in both 'classical' and 'new' ~~hosts. This can be seen qualitatively from the phylogeny~~ host populations. The virus phylogeny does hint at this result (Figure 1) but the ABC approach allows us to quantify ~~it~~ the pattern and to show that assortativity ~~might~~ may be higher for ~~the~~ 'classical' hosts. The second ~~strong~~ main result has to do with the ~~massive~~ striking difference in doubling ~~time of the epidemics between the two host types~~times. Indeed, the current spread of the epidemics in 'new' hosts appears to be at least comparable to the spread in the 'classical' hosts in the early 1990s before the advent of the third generation tests. That the duration of the infectious period in ~~new~~ 'new' hosts is in the same order of magnitude as the time until treatment suggests that the majority of the ~~infection~~ transmission events may be occurring during the acute phase. This underlines the necessity to act rapidly upon detection, for instance ~~with~~ by emphasising the importance of protection measures ~~(condom use~~

~~) and treatment initiation~~ such as condom use and by initiating treatment even during the acute phase [?]. A better understanding of the underlying contact networks ~~therefore seems essential~~ could provide additional information regarding the structure of the epidemics and, with that respect, next generation sequence data could be particularly informative [?, ?, ?].

~~Two legitimate interrogations about the study have to do with~~ Some potential limitations of the study are related to the sampling scheme~~and~~ , the assessment of the host type, and the transmission model. Regarding the sampling, the proportion of ~~the infections in~~ infected 'new' host that are sampled is ~~estimated to~~ unknown but could be high. For the 'classical' hosts, we selected a representative subset of the patients detected in the area ~~. Regarding the host type~~but this sampling is likely to be low. However, the effect of underestimating sampling for the new epidemics would be to underestimate its spread, which is already faster than the classical epidemics. In general, implementing a more realistic sampling scheme in the model would be possible but it would require a more detailed model and more data to avoid identifiability issues. Regarding assignment of hosts to one of the two types, this was ~~assessed~~ performed by clinicians independently of the sequence data. The main criterion used was the infection stage (acute or chronic), which was complemented by other epidemiological criteria (history of intravenous drug use, blood transfusion, HIV status). Finally, the 'classical' and the 'new' epidemics appear to be spreading on contact networks with different structures. However, such differences are beyond the level of details of the birth-death model we use here, and would require a larger dataset for them to be inferred.

In order to test whether the infection stage (acute vs.~~chronic) might not~~ chronic) can explain the data ~~as well~~better than the existence of two host types, we developed an alternative model where all infected hosts first go through ~~the~~ an acute phase before recovering or progressing to the chronic phase. As for the model with two host types, we used 3 time intervals. Interestingly, it was almost impossible to simulate phylogenies ~~under this model. This is most likely due to the fact that there cannot be an assortativity parameter in this alternative model (all new infections must be acute), which makes it more difficult to reproduce the observed phylogeny~~with this model, most likely because of its intrinsic constrains on assortativity (both acute and chronic infections always generate new acute infections).

~~The phylodynamics analysis raised technical challenges because of the known heterogeneity in the host population.~~ To our knowledge, ~~only two studies have recently tackled this issue~~ few attempts have been made in phylodynamics to tackle the issue of host population heterogeneity. In 2018, a ~~group~~ study used the structured coalescent model ~~,~~ to investigate the importance of accounting for so-called 'superspreaders' in the recent ebola epidemics in West Africa ~~[?]~~ [?]. The same year, another ~~group~~ study used the birth-death model to study the effect of drug resistance mutations on the R_0 of HIV strains [?]. Both of these are implemented in Beast2. However,the birth-death model is unlikely to be directly applicable to our HCV epidemics because it links the two epidemics via mutation (a host of type A becomes a host of type B), whereas in our case the linking is done via transmission (a host of type A infects a host of type B).

~~This study shows that the~~ Overall, we show that our ABC approach, which ~~had been~~ we validated for simple epidemiological models such as Susceptible-Infected-Recovered [?]~~can also~~ , can be applied to more elaborate models that ~~most~~ current phylodynamics methods have difficulties

to ~~include~~capture. Further increasing the level of details in the model ~~will require further analyses. This~~ may require to increase the number of simulations but also to introduce new summary statistics. Another promising perspective would be to combine sequence and incidence data. Although this could not be done here due to the limited sampling, such ~~a combination of different sources of data can be readily performed in an ABC framework~~data integration can readily be done with regression-ABC.

Material and methods

Epidemiological data

The Dat'AIDS cohort is a collaborative network of 23 French HIV treatment centers covering approximately 25% of HIV-infected patients followed in France (Clinicaltrials.gov ref NCT02898987). The epidemiology of HCV infection in the cohort has been extensively described from 2000 to 2016 [**?**,**?**,**?**]. The incidence of acute HCV infection has been estimated among HIV-infected MSM between 2012 and 2016~~and~~ , among HIV-negative MSM enrolled in PrEP between in 2016-2017 [**?**] ~~. The epidemiology of acute HCV infection, including incidence estimates, in~~ and among HIV-infected and HIV-negative MSMs ~~has been described~~ from 2014 to 2017 [**?**]. [SA: A réécrire pour ne citer que les données de séquences q

HCV ~~sequences~~ sequence data

We included HCV molecular sequences of all MSM patients ~~($N = 68$)~~ diagnosed with acute HCV genotype 1a infection at the Infectious Disease Department of the Hospices Civils de Lyon, France, and for whom NS5B sequencing was performed between January 2014 and December 2017 ~~were considered~~($N = 68$). HCV genotype 1a isolated from $N = 145$ non-MSM, HIV-negative, male patients of similar age were analysed by NS5B sequencing at the same time for phylogenetic analysis. This study was conducted in accordance with French ethics regulations. All patients gave their written informed consent to allow the use of their personal clinical data. The study was approved by the Ethics Committee of Hospices Civils de Lyon.

HCV testing and sequencing

HCV RNA was detected and quantified using the Abbott RealTime HCV assay (Abbott Molecular, Rungis, France). The NS5B fragment of HCV was amplified between nucleotides 8256 and 8644 by RT-PCR as previously described and sequenced using the Sanger method. Electrophoresis and data collection were performed on a GenomeLab^{TM} GeXP Genetic Analyzer (Beckman Coulter). Consensus sequences were assembled and analysed using the GenomeLab^{TM} sequence analysis software. The genotype of each sample was determined by comparing its sequence with HCV reference sequences obtained from GenBank.

Nucleotide accession numbers

All HCV NS5B sequences isolated in MSM and non-MSM patients reported in this study were submitted to the GenBank database. The list of Genbank accession numbers for all sequences is provided in Appendix.

Dated viral phylogeny

~~We inferred a maximum likelihood phylogeny using PhyML v3.0 software complemented with the Smart Model-Selection (SMS) software, from the ATGC platform [?, ?], to perform model selection. The SMS tool selected the GTR + Γ + I model with~~ To infer the time-scaled viral phylogeny from the alignment we used a Bayesian Skyline model in BEAST v2.4.8 [?]. The general time reversible (GTR) nucleotide substitution model was used with a strict clock rate fixed at 10^{-3} based on data from Ref. [?] and a gamma distribution with four substitution rate categories. The ~~maximum likelihood phylogeny was then rooted using BEAST v2.4.8 [?]. To do so, two trees were built using~~ MCMC was run for 100 million iterations and samples were saved every 5,000 iterations. We selected the maximum clade credibility using TreeAnnotator BEAST2 ~~according to two molecular clock models: either relaxed or strict [?]. We performed a model comparison with Tracer v.1.6.0 using the AIC criterion. The strict molecular clock model had a lower AIC value and was therefore considered to be the best model. The~~ package. The date of the last common ancestor was estimated ~~at 1981.34~~ to be 1977.67 with a 95% Highest Posterior Density (HPD) of ~~[1962.03; 1997.26]~~[1960.475; 1995.957].

Epidemiological model and simulations

We ~~assumed~~ assume a Birth-Death model with two hosts types (Supplementary Figure S1) with 'classical' hosts (numbered 1) and new hosts (numbered 2). This model is described by the following system of ordinary differential equations (ODEs):

$$\frac{dI_1}{dt} = a_1\beta I_1 + (1-a_2)\nu\beta I_2 - \gamma_1 I_1 \tag{1a}$$

$$\frac{dI_2}{dt} = a_2\beta\nu I_2 + (1-a_1)\beta I_1 - \gamma_2 I_2 \tag{1b}$$

~~In this~~ In the model, transmission events are possible within each type of hosts and between the two types of hosts at a transmission rate β. ~~The parameter~~ Parameter ν corresponds to the transmission ~~differential between the number of partners of the classical hosts (I_1) and that of the new hosts(I_2). Individuals I_i~~ rate differential between classical and new hosts. Individuals can be 'removed' ~~from the infectious compartment i~~ at a rate γ_1 from an infectious compartment (I_1 or I_2) via infection clearance, host death or change in host behaviour (e.g. condom use). ~~This event occurs at a removal rate γ_1.~~ The assortativity between host types~~is given by the a_i (a value close to 1 means there is very little transmission to~~, which can be seen as the percentage of transmissions that occur with hosts from the ~~other type)~~same type, is captured by parameter a_i.

The ~~basic~~ effective reproduction number (denoted R_0) is the number of secondary cases caused by an infectious individual in a fully susceptible host population [?]. We seek to infer the R_0 from

Table 1. Prior distributions for the birth-death model parameters over the three time intervals. t_0 is the date of origin of the epidemics in the studied area, t_1 is the date of introduction of 3rd generation HCV tests, t_2 is the date of emergence of the epidemic in 'new' hosts and t_f is the time of the most recent sampled sequence.

Interval height	~~γ_1~~ γ_i	~~γ_2~~ ν	$R_0^{(1)}$	a_i
$[t_0, t_1]$	Unif$(0.1, 4)$	~~Unif$(0.1, 4)$~~ 0	Unif$(0.9, 15)$	Unif$(0, 1)$
$[t_1, t_2]$			Unif$(0.1, 3)$	~~Unif(0.13)~~
$[t_2, t_3]$		Unif$(0, 10)$	~~Unif$(1, 10)$~~	

the classical epidemic, denoted $R_0^{(1)}$ and defined by $R_0^{(1)} = \beta/\gamma_1$, ~~and~~ as well as the R_0 of the new epidemic, denoted $R_0^{(2)}$ and defined by ~~$R_0^{(2)} = \nu\beta/\gamma_2 = \nu R_0^{(1)}\gamma 1/\gamma_2$~~ $R_0^{(2)} = \nu\beta/\gamma_2 = \nu R_0^{(1)}\gamma_1/\gamma_2$.

The doubling time of an epidemics (t_D) ~~correspond~~ corresponds to the time required for the number of infected hosts to double in size~~and it~~. It is usually estimated in the early stage of an epidemics, when epidemic growth can assumed to be exponential. ~~Here, we assumed~~ To calculate it, we assume perfect assortativity ($a_1 = a_2 = 1$) and ~~approximated~~ approximate the initial exponential growth rate by $\beta - \gamma_1$ for 'classical' hosts and $\nu\beta - \gamma_2$ for 'new' hosts. Following [**?**], we obtain $t_D^{(1)} = \ln(2)/(\beta - \gamma_1)$ and $t_D^{(2)} = \ln(2)/(\nu\beta - \gamma_2)$.

~~$R_0^{(1)}$ is assumed to vary over~~ We consider three time intervals~~: from t_0 to t_1, from t_1 to t_2, from t_2 to t_f~~. During the first interval $[t_0, t_1]$, t_0 being the year of the origin of the epidemic in the area of Lyon, we assume that only classical hosts are present. The second interval $[t_1, t_2]$, begins in $t_1 = 1997.3$ with the introduction of the third generation HCV tests, which we assume to have ~~decreased~~ affected $R_0^{(1)}$ through the decrease of the transmission rate β. Finally, the 'new' hosts appear during the last interval $[t_2, t_f]$~~. We also wish to date the origin of this second outbreak(~~, where t_2~~)~~, which we infer, is the date of origin of the second outbreak. The final time (t_f) is ~~given by the sampling date of our most recent sequence, which is~~ set by the most recent sampling date in our dataset (2018.39). The prior distributions used are summarized in Table 1 and shown in Figure 2.

~~We used~~ To simulate phylogenies, we use a simulator implemented in `R` via the `Rcpp` package~~to simulate epidemiological trajectories and transmission sampled trees. The simulator resembles that developed by [?] and uses Gillespie's stochastic simulation algorithm to simulate epidemiological trajectoriesgiven our model. Further details about this simulator can be found elsewhere preprint by Danesh et al. to be submitted to bioRxiv.~~

~~Following other phylodynamics studies, we assume that a time-scaled phylogenyof an epidemic can be correlated to a sampled transmission tree in which a branching represents a transmission event and a leaf represents a~~. This is done in a two-step procedure. First, epidemiological trajectories are simulated using the compartmental model in equation 1 and Gillespie's stochastic event-driven simulation algorithm [**?**]. The number of individuals in each compartment and the reactions occurring through the simulations of trajectories, such as recovery or transmission events,

are recorded. Using the target phylogeny, we know when sampling events occur. For each simulation, each sampling date is randomly associated to a host compartment using the observed fraction of each infection type (here 68% of the dates associated with 'classical' hosts type and 32% with 'new' hosts). Once the sampling dates are added to the trajectories, we move to the second step, which involves simulating the phylogeny. This step starts from the last sampling date and follows the epidemiological trajectory through a coalescent process, that is backward-in-time. Each backward step in the trajectory can induce a tree modification: a sampling event ~~. Here, our simulator generates phylogenies of infections using the coalescent approach based on simulated trajectories and sampling dates. Importantly, we~~ leads to a labelled leaf in the phylogeny, a transmission event can lead to the coalescence of two sampled lineages or to no modification of the phylogeny (if one of the lineages is not sampled).

We implicitly assume that the sampling rate is low, which is consistent with the limited number of sequences in the dataset. We also assume that the virus can still be transmitted after sampling.

We ~~simulated 61,000~~ simulate 71,000 phylogenies from known parameter sets drawn in the prior distributions shown in Table~~1. These were~~ 1. These are used to perform the rejection step and build the regression model in the Approximate Bayesian Computation (ABC) inference.

ABC inference

Summary statistics

Phylogenies are rich objects and to compare them we ~~used~~ break them into summary statistics. These ~~were~~ are chosen to capture the epidemiological information ~~that we wanted to extract~~of interest. In particular, ~~we used the summary statistics based on~~ following an earlier study, we use summary statistics from branch lengths, ~~topology of the tree~~ tree topology, and lineage-through-time (LTT) ~~plot developed by~~ [?].

We also ~~computed additional~~ compute new summary statistics to extract information regarding the heterogeneity of the population, the assortativity, and the difference between the two R_0. To do so, we ~~annotated~~ annotate each internal node by associating it with a probability ~~of being~~ to be in a particular state (here the ~~type of host , classical or new). This probability was assumed to be~~ host type, 'classical' or 'new'). We assume that this probability is given by the ratio

$$P(Y) = \frac{\text{number of leaves labelled } Y}{\text{number of descendent leaves}} \tag{2}$$

where Y is a ~~type of host .~~

~~Each node could therefore be~~ state (or host type). Each node is therefore annotated with n ratios, n being the number of possible states~~(i. e. types of label).~~ . Since in our case $n = 2$, we only ~~followed~~ follow one of the labels and ~~used~~ use the mean and the variance of the distribution of the ratios (one for each node) as summary statistics.

In a phylogeny, ~~'cherries'~~ cherries are pairs of leaves that are adjacent to a common ancestor. There are $n(n+1)/2$ categories of cherries. Here, we ~~counted the number~~ compute the proportion of homogeneous cherries for each label and the ~~number~~ proportion of heterogeneous

cherries. ~~Furthermore, we considered triplets, that is~~ We also consider pitchforks, which we define as a cherry and a leaf adjacent to a common ancestor, and ~~introduced~~ introduce three categories: homogeneous ~~triplets, triplets whose cherries were~~ pitchforks, pitchforks whose cherries are homogeneous for a label and whose leaf ~~was~~ is labelled with another trait, and ~~triplets whose cherries were heterogeneous. We expected the structure of cherries and triplets capture the information about the interaction between the different hosts.~~ pitchforks whose cherries are heterogeneous.

The Lineage-Through-Time (LTT) plot displays the number of lineages of a phylogeny over time~~:~~. In this plot, the number of lineages is incremented by one ~~for each~~ every time there is a new branch in the phylogeny, and is decreased by one ~~for each~~ every time there is a new leaf in the phylogeny. We ~~used~~ use the ratios defined for each internal node to build ~~an~~ a LTT for each label type, which we refer to as ~~an~~ 'LTT label plot'. After each branching event in phylogeny, we ~~incremented~~ increment the number of lineages by the value of the ratio of the internal node for the given label. This number of lineages ~~was~~ is decreased by one ~~for each~~ every time there is a leaf in the phylogeny. In the end, we ~~obtained~~ obtain $n = 2$ LTT label plots.

Finally, for each label, we ~~computed some of the same~~ compute some of our branch lengths summary statistics ~~as for the unlabelled phylogeny on homogeneous clusters and heterogeneous clusters~~ on homogeneous clades and heterogeneous clades present in the phylogeny. Homogeneous ~~clusters were~~ clades are defined by their root having a ratio of 1 for one type of label and their size being greater than $N_{\min}$. For heterogeneous ~~cluster, we kept~~ clades, we keep the size criterion and ~~imposed~~ impose that the ratio ~~was~~ is smaller than 1 but greater than a threshold ϵ. After preliminary analyses, we set $N_{\min} = 4$ leaves and $\epsilon = 0.7$. We therefore ~~obtained~~ obtain a set of homogeneous ~~clusters~~ clades and a set of heterogeneous ~~clusters~~ clades, the branch lengths of which ~~were pooled~~ we pool into two sets to compute the summary statistics of heterogeneous and homogeneous ~~clusters.~~ clades. Note that we always select the largest clade, for both homogeneous and heterogeneous cases, to avoid redundancy.

Regression-ABC

We first ~~measured~~ measure multicollinearity between summary statistics using variance inflation factors (VIF). Each summary statistic ~~was~~ is kept if its VIF value ~~was~~ is lower than 10. This ~~step led~~ stepwise VIF test leads to the selection of 88 summary statistics out of 234.

We then ~~used~~ use the **abc** function from the `abc` `R` package to infer posterior distributions ~~from rejection only~~ generated using only the rejection step. Finally, we ~~performed~~ perform linear adjustment using an elastic net regression.

The ~~abc~~ **abc** function performs a classical one-step rejection algorithm [**?**] using a tolerance parameter P_δ, which represents a percentile of the simulations that are close to the target. ~~For this, we calculate a Euclidian distance between the simulations~~ To compute the distance between a simulation and the target~~using the~~, we use the Euclidian distance between normalized simulated vector of summary statistics and the normalized target vector.

Prior to linear adjustment, the `abc` function performs smooth weighting using an Epanechnikov kernel [**?**]. Then, using the `glmnet` package in `R`, we ~~implemented~~ implement an elastic-net (EN) adjustment, which balances the Ridge and the LASSO regression penalties [**?**]. The EN performing

a linear regression, ~~is it~~ it is not subject to the risk of over-fitting that may occur for non-linear regressions (e.g. when using neural networks, support vector machines or random forests).

~~We inferred posterior distributions of~~ In the end, we obtain posterior distributions for t_0, t_2, a_1, a_2, ν, γ_1, γ_2, $R_0^{(1),t_1}$ and $R_0^{(1),t_2}$ using our ABC-EN regression model with $P_\delta = 0.1$.

Parametric bootstrap and cross validation

~~Parametric bootstrap validation consisted in simulating 1,000 transmission additional trees~~

Our parametric bootstrap validation consists in simulating 5,000 additional phylogenies from parameter sets drawn in posterior distributions. We then ~~computed~~ compute summary statistics and ~~performed~~ perform a principal component analysis (PCA) on the vectors of summary statistics for the simulated and for the target data. If the posterior distribution is informative, we expect the target data to be similar to the simulated phylogenies. On the contrary, if the posterior distribution can generate phylogenies with a variety of shapes, the target data can be outside the cloud of simulated phylogenies in the PCA.

In order to assess the robustness of our ABC-EN method to infer epidemiological parameters of our BD model, we ~~performed~~ also perform a 'leave-one-out' cross-validation ~~. This consisted~~ as in [?]. This consists in inferring posterior distributions of the parameters from one simulated ~~tree~~phylogeny, assumed to be the target ~~tree~~phylogeny, using the ABC-EN method with the remaining ~~60,000 simulated trees. We ran~~ 60,999 simulated phylogenies. We run the cross-validation 100 times with 100 different target ~~trees and measured the~~ phylogenies. We consider three parameter distributions θ: the prior distribution, the prior distribution reduced by the feasibility of the simulations and the ABC inferred posterior distribution. For each of these parameter distributions, we measure the median and compute, for each simulation scenario, the mean relative error ~~of inference~~ (MRE) such as:

$$MRE = \frac{1}{100}\sum_{i=1}^{100} \mid \frac{\theta_i}{\Theta} - 1 \mid \tag{3}$$

where Θ is the true value.

Acknowledgments

We thank Jūlija Pečerska for her help with Beast2. GD is funded by the Fondation pour la Recherche Médicale (FRM grant number ECO20170637560). GD and SA acknowledge further support from the CNRS, the IRD and the itrop HPC (South Green Platform) at IRD montpellier, which provided HPC resources that contributed to the results reported here (`https://bioinfo.ird.fr/`).