Molecular epidemiology of hepatitis B in the Netherlands
van Houdt, R.

Citation for published version (APA):
van Houdt, R. (2009). Molecular epidemiology of hepatitis B in the Netherlands
Molecular sequence data of hepatitis B virus reveals sudden decrease in genetic diversity after vaccination

W. Marijn van Ballegooijen¹, Robin van Houdt², Sylvia M. Bruisten²,³, Hein J. Boot¹, Roel A. Coutinho¹,³, Jacco Wallinga¹,⁴

¹ National Institute for Public Health and the Environment, Centre for Infectious Disease Control, Bilthoven, the Netherlands
² GGD, Public Health Service, Department of Infectious Diseases, Amsterdam, the Netherlands
³ Amsterdam Medical Center, University of Amsterdam, Department of Human Retrovirology, Amsterdam, the Netherlands
⁴ Julius Center for Health Sciences and Primary Care University Medical Center Utrecht, the Netherlands

Submitted for publication.
Abstract

The effect of vaccination programs on transmission of infection is usually assessed by monitoring programs that rely on notifications of symptomatic infections. For monitoring infections with a high proportion of asymptomatic cases or a low reporting rate, molecular sequence data combined with modern coalescent-based techniques offer a complementary tool to assess the transmission of infection. Here, the authors investigate the added value of using viral sequence data for monitoring a vaccination program that was started in 1998 and targeted against hepatitis B virus among men who have sex with men in Amsterdam. The incidence in this target group, as estimated from the notifications of acute infections with hepatitis B, was low and therefore there was insufficient power to show a significant change in incidence. In contrast, the genetic diversity as estimated from the viral sequence collected among the target group revealed a marked decrease after the vaccination was introduced. Taken together, the findings suggest that introduction of vaccination coincided with a change in the target group towards behavior with higher risk of infection. The authors argue that molecular sequence data provides a powerful additional monitoring instrument, next to conventional case registration, for assessing the impact of vaccination.
Introduction

Vaccination programs are commonly monitored using clinical case registry data. Molecular sequences may offer an additional data source for assessing the impact of vaccination. Recent advances in the analysis of molecular data allow the estimation of changes in the circulation of infectious disease based on DNA or RNA sequences [1-3]. Bayesian coalescent-based methods have been used to estimate changes in the circulation of Hepatitis C Virus [4], hepatitis A virus [5], West Nile virus [6] and HIV-1 [7]. The methods assume that only a small fraction of cases per unit time are sampled and sequenced to infer the infection dynamics in the larger population. That makes them well suited for monitoring infection dynamics where conventional case notifications offer an incomplete picture because part of the infections are asymptomatic, subclinical or simply not reported. There are, however, no studies so far that have used molecular sequence data to assess the impact of a vaccination program.

A direct motivation for this study was provided by the assessment of the impact of a vaccination program against hepatitis B virus (HBV) targeted at men who have sex with men (MSM) in Amsterdam. Sexual transmission is the most frequent route of infection in the Netherlands [8-10]. An acute HBV infection typically lasts three to four months, but can lead to chronic carriage of the virus. Chronic infection is uncommon in sexually acquired HBV because the probability of developing chronic carriage declines with age, from 90% in infants to 5% in adults [11, 12]. Symptoms of acute HBV infection may include jaundice and fatigue, but often infection is asymptomatic or subclinical. The fraction of asymptomatic infections in adults is estimated to be around two thirds. Because of this high proportion of asymptomatic and subclinical cases, a significant fraction of all infections are missed by routine passive surveillance by reporting of symptomatic cases.

Infection with HBV can be prevented by vaccination. Efficacy of vaccination against infection in MSM is around 80-87% [13]. Because of the low incidence in the general population in the Netherlands, the Dutch government opted for a vaccination program targeted at risk groups. A pilot vaccination program was started in Amsterdam in 1998, aimed at MSM, commercial sex workers, heterosexuals with a high rate of partner change and drug users. Although the targeted vaccination program is under close scrutiny, the high proportion of asymptomatic and subclinical cases makes it difficult to assess the impact of vaccination and draw firm conclusions that are helpful for policy making.

Coalescent-based analysis of molecular sequences provides a possibility to overcome the difficulties in assessing the impact of targeted vaccination programs. As the coalescent-based method relies on the genealogical relationships between the sampled sequences to infer underlying changes in the population dynamics of the pathogens, its estimates are unaffected by a high proportion of asymptomatic or subclinical cases or a low notification rate. This makes it a promising complementary tool for inferring the impact of vaccination on transmission. In this study we assess the additional value of coalescent-based analysis of molecular sequence data of hepatitis B virus, next to notification of acute hepatitis B infections, in monitoring the impact of a targeted vaccination program.
Materials and Methods

Study population and vaccination program

We focus on hepatitis B infection among men having sex with men (MSM) in Amsterdam over the period 1992-2006. Of all reported acute HBV cases in Amsterdam, half were reported among MSM [10]. The total size of the MSM risk group in Amsterdam is loosely estimated at 26,000 persons [14]. In this risk group genotype A accounts for 87% of all infections, and almost 84% of genotype A infections are among MSM [10]. Phylogenetic trees of HBV sequences from all genotypes from the study population and other risk groups in the area (up to 2003) can be found in van Houdt et al. 2006 [10]. The targeted vaccination program was initiated in 1998. The vaccination effort for MSM consisted of an outreach program for recruiting vaccinees at social venues and offering vaccination for free. Over the period from 1998 to 2006, 8,472 persons took the first vaccination. Of those taking the first vaccination, 1,597 were found to be already anti-HBc positive, indicating prior exposure, and were not further vaccinated. We note for completeness that the persons vaccinated were not randomly sampled from the risk group, so that translating this figure into seroprevalence is difficult. A completed vaccination regimen of three administrations was reached by 4,479 persons.

Data collection: notifications of acute infections with hepatitis B

Observed cases of hepatitis B in Amsterdam are reported to the Municipal Health Service. The patients are approached by public health nurses who collect general information such as age and gender and who try to identify the most likely route of transmission i.e. homosexual or heterosexual contact, injecting drug use, parent to child, etc. For some cases a transmission route could not be established. Over the period from 1992 to 2006 there were 186 reported cases in Amsterdam of acute hepatitis B infections with a reported homosexual transmission route. In the same period, there were 118 reported cases attributed to heterosexual transmission, 56 from unknown source and 53 from various non-sexual sources.

Analysis of notified acute infections with hepatitis B

We are interested in observing a change in incidence of infection. We use change point analysis to detect whether there is any empirical support for a change, and if so, when this change occurred. For such an analysis we invoke a step function to describe incidence of infection. This function requires three variables: the incidence before the change, the moment of change, and the incidence after the change. We used all registered clinical cases that were attributed to homosexual transmission based on interviews and assumed a constant notification probability (see appendix for technical details). We use a Bayesian analysis to assess the posterior probability of many possible stepwise functions in incidence, given the data. The average over all these stepwise functions, each weighted by their respective posterior probability, yields an estimated gradual change in incidence of infection as well as a probability distribution for the moment of change in incidence of infection. We
note that a Bayesian change-point analysis does not assume that a step function is the most appropriate model for change, but that it assumes that the step function is an appropriate building block for an appropriate model of change. The use of change point analysis to detect the signature of different shapes of changes is of special value when the exact shape of change is not known beforehand. We use this method to examine indications of a change in incidence over the period 1992-2003 that was also analyzed by van Houdt et al. [10], see also figure 1 for the data.

\textit{Data collection: DNA sequences of hepatitis B virus}

For our analysis, we took all 54 available sequences of HBV genotype A. Of these, we excluded all 3 sequences that were collected from female patients and all 8 sequences that were not self-reported to originate from homosexual transmission. Additional analysis showed that exclusion of these 11 sequences did affect the precision, but not the values of resulting estimates. Available sequences were obtained and aligned as described earlier [10], see the appendix for accession numbers. The 43 sequences are 654 nt. DNA fragments of the S gene, including part of the pre-S2 region. Collection dates for these sequences were spread out evenly over the study period from 1992 to 2006, leading to inclusion of approximately three sequences per year.

\textit{Analysis of molecular DNA sequences}

Coalescent-based methods relate the coalescence (that is, inverse branching) rate $\theta$ of a phylogenetic tree to the so-called effective number of infections $N_e$ according to $\theta = c \sigma^2 / (\tau N_e)$. In this equation the parameter $c$ corrects for the number of branches coexisting at a particular time in the phylogenetic tree and $\sigma^2$ is the variance in the number of new infections that infected cases make [15,16]. The generation time $\tau$ stands for the typical time between subsequent infections. The generation time $\tau$ and variance $\sigma^2$ are not well known for many pathogens --- for sexually transmitted HBV in MSM, we assume the generation time to lie between two and four months -- and for this reason $N_e \tau / \sigma^2$ is commonly expressed as a compound variable. Following standard terminology, we refer to this compound variable $N_e \tau / \sigma^2$ as genetic diversity [e.g. 17,18].

We use a Bayesian skyline plot [19] to detect trends in the transmission of HBV. The procedure to construct these plots is implemented in the BEAST package [2]. Skyline plots present a flexible model to visualize the trends in population change suggested by the data, making few prior assumptions about such changes.

We are interested in observing a change in the compound variable genetic diversity. We use change point analysis to detect whether there is any empirical support for a change, and if so, when this change occurred. A step function was used to describe genetic diversity. This requires three variables: genetic diversity before the change, the moment of change, and genetic diversity after the change. We compare this change point model with a model based on a constant number of infections over the entire time period (representing the null hypothesis of no change in HBV circulation). To describe the goodness-of-fit of both models we used the marginal likelihood, using the smoothed method of Newton & Raftery [20] as
implemented in the BEAST package (see appendix for technical details). To quantify the strength of statistical evidence in the molecular sequence data for a change in prevalence we used Bayes Factors [21]; the interpretation of these Bayes Factors is similar to the more familiar likelihood ratio.

Results

The clinical case registries for the MSM in the Amsterdam area, going back to 1992, provide no support for a change in number of reported acute HBV infections (figure 1). Accordingly, the posterior density from the change point analysis is spread out over the studied period, indicating no clear-cut moment at which a change in incidence might have occurred (figure 2). Such a change-point would become apparent in figure 2 as a concentration of the posterior probability density in a limited period in time.

The maximum posterior density phylogenetic tree of the Bayesian Skyline analysis is presented in figure 3. The sequence data reveal a decline in the genetic diversity (figure 4) for which there is strong statistical support (table 1). Log marginal likelihoods are used to indicate the goodness-of-fit (the higher the better) of each model. The log marginal likelihoods were approximated by using the smoothed Newton & Raftery (15) method. Bayes factors indicate the strength of evidence for one model over another, in a similar way as the likelihood ratio. The common guidelines for interpreting log Bayes factors (and log likelihood ratios) is that values of 0-1/2 provide no evidence, values of 1/2-1 provide weak evidence, values of 1 and up provide strong to decisive evidence in favor of the hypothesis with the maximum marginal likelihood [21]. Here the log marginal likelihood indicates that the model with a change in HBV genetic diversity gives the best fit to the data; the log Bayes

Figure 1. Annual number of registered acute HBV cases in the Amsterdam area. Data is categorized according to probable source of infection: Homosexual transmission (MSM, solid), heterosexual transmission (dotted) and unknown transmission (dashed). The numbers of registered cases for MSM show little indication of a change after introduction of vaccination in 1998.
Figure 2. Dating a moment of change in Hepatitis B genotype A incidence among MSM from clinical case registries. The case reports from the 1992-2003 period were translated into a Bayesian posterior probability estimate for the true number of cases based on the assumption that actual cases are notified with a probability of 20%. The clinical case registries show no apparent date of change in the number of HBV cases.

Figure 3. Maximum clade credibility phylogenetic tree from the Bayesian Skyline analysis. Information from the time spaced samples is used to calibrate the phylogeny in real time.
Figure 4. Trends in the genetic diversity of Hepatitis B (genotype A) in MSM in Amsterdam, the Netherlands. This Bayesian Skyline plot (gray area) was estimated from viral sequences collected between 1992 and 2006. The genetic diversity is expressed as $N_e \tau / \sigma^2$, where $N_e$ is the effective number of infections, $\tau$ is the generation interval and $\sigma^2$ is the variance in the number of new infections that infected cases make. The median genetic diversity declines sharply during the period 1998-2000 (black line).

Table 1. Comparison of the Hypothesis of a Constant HBV Genetic Diversity (Model 1) With the Hypothesis of a Change in Genetic Diversity (Model 2).

<table>
<thead>
<tr>
<th></th>
<th>Model 1</th>
<th>Model 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>HBV genetic diversity</td>
<td>Constant</td>
<td>Stepwise change</td>
</tr>
<tr>
<td>Mean genetic diversity</td>
<td>308</td>
<td>602</td>
</tr>
<tr>
<td>($N_e \tau / \sigma^2$)</td>
<td>97</td>
<td>126-1417</td>
</tr>
<tr>
<td>95% Credible interval</td>
<td>68-996</td>
<td>5-290</td>
</tr>
<tr>
<td>$10^{10}\log$ Marginal likelihood (L)</td>
<td>-1332.1</td>
<td>-1329.1</td>
</tr>
<tr>
<td>$10^{10}\log$ Bayes factor (Lmax-L)</td>
<td>3.0</td>
<td>0</td>
</tr>
</tbody>
</table>

factor indicates the data provides strong evidence for a model with a change in HBV genetic diversity over a model with a constant genetic diversity.

The population histories sampled from the posterior shows a mean reduction of the genetic diversity to 17% (95% CI is 3%, 65%) of its former magnitude. The interval estimate of the relative change is below 100% and this provides therefore strong statistical support for the change being a decrease. The stepwise model places a high posterior probability on a change occurring just after the implementation of the vaccination program in 1998, with a maximum posterior probability near the year 2000 (figure 5).

Coalescent based approaches can underestimate genetic diversity when the number of lineages in the phylogenetic tree is large compared to the number of infected persons in the
population. This can be cause for concern when in a phylogeny the number of lineages increases toward the tips of the tree. Because of the distribution of our samples over time, however, change in the number of lineages through time throughout the study period is relatively low. Visual inspection of lineage through time plots from the posterior sample of trees indicates the number of lineages ranges roughly between 3 and 15 between 1992 and 2003, with no indication of an increasing or decreasing trend (data not shown). Fu et al. [22] show that the Kingman coalescent can be a reasonable approximation to exact coalescent results even for small and oversampled populations.

**Figure 5.** Dating the moment of change in Hepatitis B genotype A genetic diversity among MSM from sequence data. The graph represents the posterior probability distribution for this date in the stepwise model, given the sequence data. The stepwise model we used allows for a single change in genetic diversity between 1991 and 2003. The maximum posterior probability density, peaking around 2000.

**Discussion**

In this study we explored the added value of using molecular sequence data, next to clinical case registries, in assessing the impact of vaccination. As an example, we used a targeted vaccination program against hepatitis B infections among MSM in Amsterdam, because infections with hepatitis B are often asymptomatic and will not be notified. We observed a significant decline in genetic diversity of the hepatitis B genotype A viral sequences collected among the target population, and this decline occurred within a few years after implementation of the vaccination program. In contrast, we could not detect a significant decline in number of notified acute hepatitis B infection in the target population. Apparently, we have two opposing observations and it may be tempting to argue that one of them is flawed. Both data sources, however, may have limitations. On the one hand, since we have included only 43 sequences (645 nucleotides) our number of observations may not be large enough and our molecular sequence data may contain insufficient information on the actual incidence of hepatitis B among MSM in Amsterdam. We have collected hepatitis
B viral sequences of the S gene, and mutations in the S gene are constrained by the overlapping reading frame of the P gene. Inclusion of sequences from the more variable C gene could potentially further enhance the information that can be obtained from such sample sets [23]. But despite the small number of sequences and the constrained mutation rate, we do pick up a statistically significant signal in the data. The support of the data for a change in diversity is classified as strong evidence by all possible classifications we have consulted. There is a small probability that we have a type I error and reject the hypothesis that transmission remained constant while this hypothesis was true. With smaller sample sizes this probability becomes smaller; in our case this means that small sample size cannot be the reason for detecting a change. This rules out the possibility that transmission remained constant and we have detected an erratic signal due to a small number of sequences; the limited number of sequences do carry a definite signal that transmission has changed.

On the other hand, the case notification data contain insufficient information about incidence, due to incomplete case ascertainment of all acute infections among MSM. We have only 186 notified acute infections of hepatitis B over a period of 15 years. Some infections through homosexual contacts are suspected to be reported as unknown. Case reports of acute infections through heterosexual contacts and acute infections with unknown route declined after 1998 [10]. Moreover, it cannot be excluded that increased awareness due to the vaccination program in both the risk group and healthcare workers have influenced reporting practice. These factors suggest that it is difficult to infer the absolute value of incidence from case notifications. There is a small probability that we have a type II error and accept the hypothesis that transmission remained constant while there was a change. For larger changes, this possibility diminishes. It is therefore extremely unlikely that a large change would be missed; the limited number of case reports does signify that incidence has changed very little. Case reports for IDU and heterosexual risk groups (both with smaller numbers of reported cases then MSM in the pre-vaccination period) did decline significantly in the 6 years after vaccination [10], indicating that a successful vaccination program can cause a detectable decline in case reports on this timescale and group size.

Most studies that relate molecular sequence data to infection dynamics assume proportionality between the genetic diversity and incidence of infection. However, our results demonstrate that this relation might be more complicated than suggested by the standard assumption of proportionality. We can offer a number of non mutually exclusive explanations for the apparent discrepancy between genetic diversity and incidence. A reduction in genetic diversity can be due to a decline in number of infections but also to a shorter generation time or an increase in the average and variance of the number of secondary infections produced by one infective individual [14]. The latter could occur if the vaccination program coincided with an increase in risk behavior. If risk behavior increased among MSM in Amsterdam, we would also expect to see an increase in incidence of other sexually transmitted infections in this group. This is indeed observed [24-26]. Improved HIV-treatment is thought to have led to increased risk behavior around 1995 - 2000 in the Amsterdam region and internationally [24-28]. Therefore it is very likely that there has been a change towards increased risk
behavior among the target population during the time of introduction of vaccination. An increase in super spreaders (other then HIV+ MSM) can also help to explain reduced diversity. Having these coincidental changes, we would expect that incidence remains about constant, as increased risk behavior would result in increased incidence, and as increased vaccination uptake would result in lower incidence, and we would expect genetic diversity to decrease sharply, as both increased risk behavior and increased vaccination uptake would result in lower genetic diversity. Xiridou et al. [29] show in a modelling study that that negative effect of increased risk behavior among MSM can counterbalance the positive of HBV vaccination at current levels of vaccine uptake in Amsterdam. Furthermore, increased risk behavior from chronic carriers of HBV, including HIV-positive persons that can carry high viral loads, could have introduced conserved HBV strains into the MSM population, leading to a further decline of genetic diversity observed in sampled sequences. These behavioral changes can explain why the decline in genetic diversity is larger and more sudden than the expected decline in genetic diversity that would result from a decline in number of infections. Finally, imported strains can potentially confound our dataset and increase the estimated genetic diversity in the period before their introduction. A standard coalescent based analysis, without accounting for case notifications, might risk interpreting the resulting loss of genetic diversity in samples as an additional decline in the effective number of infections $N_e$. A standard analysis of case notifications, without accounting for sequence data, might miss the changes in transmission altogether.

Summarizing, after the introduction of a targeted hepatitis B vaccination program we did not observe a distinct change in notified acute infections, whereas we did observe a significant sharp decline in genetic diversity in the hepatitis B genotype A that is characteristic for the target population. The molecular data reveal that HBV transmission started changing immediately after the implementation of the vaccination program. A plausible explanation for these observations is that the introduction of the program coincided with an increased risk behavior, leading to reduced diversity in the sampled sequences. Studies based only on information about case notifications would have missed the change in genetic diversity altogether; whereas studies based only on information about molecular sequence data would have overestimated the size of the impact of the targeted vaccination program. We conclude that the use of molecular sequence data is a valuable addition to routine case notifications, both for assessing and interpreting the impact of vaccination on transmission of the pathogen.

References


Appendix

Change point analysis of registered acute HBV infections

We assumed that the registered clinical cases \( C_t \) in year \( t \) represent a binomial sample from the true incidence in that year \( \lambda_t \), based on a constant notification probability \( p \). Similar to the model used in the molecular analysis, we assume incidence is constant over time with the exception of a single stepwise change at time \( T \). We use flat priors for the incidence before and after this stepwise change over the range from 1 to 500 persons per year. We assume that the prior probability of a change in number of infections at time \( T \) is constant. We use a
constant prior probability for number of infections after change, \( I_a \), and number of infections before change, \( I_b \). With this model we examine the posterior probability for the change occurring in a particular year, given the clinical case registries. Using Bayes theorem, the marginal posterior for the change occurring in year \( T \) is

\[
P(T \mid C_{1992}, \ldots, C_{2003}) = \frac{\sum_{I_a} \sum_{I_b} \prod_{t=1992}^{T} \text{Bin}(C_t \mid I_b, p) \prod_{t=T+1}^{2003} \text{Bin}(C_t \mid I_a, p)}{\sum_{T} \sum_{I_a} \sum_{I_b} \prod_{t=1992}^{T} \text{Bin}(C_t \mid I_b, p) \prod_{t=T+1}^{2003} \text{Bin}(C_t \mid I_a, p)}.
\]

**Change point analysis of HBV DNA sequences**

Our dataset of 43 sequences contained 21 different strains; and the dominant strain contained 21 sequences. The Bayesian framework is appropriate to such datasets because it can deal with identical sequences without reverting to branches of length zero. We used the HKY+I model [30] for molecular evolution in all analyses presented. To prevent overfitting we constructed the skyline plots from a model with three constant levels of population size and two points where this level changed. BEAST uses Markov Chain Monte Carlo (MCMC) sampling to estimate the posterior distribution for parameters of interest. The MCMC chain was run for 88 million updates, results were visually inspected for convergence and the first 8 million updates were discarded as burn-in.

We calculated posterior distributions for the parameter values of interest (time of change, genetic diversity before change, genetic diversity after change). We use an uninformative, flat prior for the time of change within the years 1991-2003 and we used flat priors ranging from 0 to 2000 for the product of the generation time and the genetic diversity before and after this change. Because sequence data only contain information on genetic diversity in the population at a time before the sequences were collected, and because the collection dates were spread out over the study period, the dataset contains limited information on the most recent years. The limitation of the changepoint window to 2003 was made to prevent a strong influence of the Bayesian prior in the recent years on the estimate. We examined this changepoint model running the MCMC chain for 88 million steps examining for convergence by visual inspection and discarded the first 8 million steps as burn-in.

**Tests on simulated data**

To check the reliability of our method we redid our analysis on simulated datasets. First, to show that the inferred decline in genetic diversity depends crucially on information in the observed sequences and their associated dates, we re-assigned collection dates to the sequences at random, by sampling without replacement from the actual collection dates of the sequences. The results of Bayesian skyline analysis on 20 randomized datasets (figure 6) show that the Bayesian skyline plot from the actual data is distinct from the skyline plots of sequences with randomized dates. This indicates that the observed pattern is not
determined by the method of analysis or the sampling scheme, but follows from the actual
dates at which the sequence samples were collected.
Second, to show that a decline in genetic diversity is unlikely for a population in which the
population genetic parameters are constant through time, we simulated molecular evolution
on 100 random coalescent phylogenies under the assumption of constancy, and analysed
these. We used the mean posterior estimates of coalescent and evolutionary parameters
from the constant population model of the HBV sequences. The random coalescent trees
used the same sampling timepoints as our data. Molecular sequence evolution was
simulated on these trees using the Mesquite program [31]. Change point analysis of these
simulated datasets shows that their typical posterior change point distribution is
uninformative (roughly uniform throughout its interval, figure 7). A decline in genetic diversity
was not apparent in the simulated datasets. The posterior distributions of the change point
for the simulated datasets do not show a concentration of probability density in particular
time intervals, unlike the posterior distribution for the actual dataset. This indicates that the
decay in genetic diversity observed in the HBV data is unlikely to be originated under the null
hypothesis of no change in genetic diversity. Further tests of the coalescent method
implemented in BEAST on simulated data can be found in Drummond et al. [2].

Figure 6. Trends in the genetic diversity in artificial datasets with resampled dates. The
magnitude and pattern of mean Bayesian Skyline estimates for the artificial data (gray lines)
differ considerably from the mean Bayesian Skyline for the original dataset (black line).
Figure 7. Posterior density of change-point analysis for simulated data (gray lines show 100 replicates). Dashed black lines denote the 90% envelope. Solid black line represents the change-point posterior from the original data for comparison.

Sequence accession numbers