Virus Evolution, 2017, 3(2): vex028

doi: 10.1093/ve/vex028
Research article

Accurate quantification of within- and between-host HBV evolutionary rates requires explicit transmission chain modelling

Bram Vrancken,[1,*,†] Marc A. Suchard,[2,3,4] and Philippe Lemey[1,‡]

[1]Department of Microbiology and Immunology, Rega Institute for Medical Research, KU Leuven – University of Leuven, B-3000 Leuven, Belgium, [2]Department of Biomathematics, University of California, Los Angeles, CA 90095, USA, [3]Department of Human Genetics, David Geffen School of Medicine at UCLA, University of California, Los Angeles, CA 90095, USA and [4]Department of Biostatistics, UCLA Fielding School of Public Health, University of California, Los Angeles, CA 90095, USA

*Corresponding author: E-mail: bram.vrancken@kuleuven.be

[†]http://orcid.org/0000-0001-6547-5283

[‡]http://orcid.org/0000-0003-2826-5353

Abstract

Analyses of virus evolution in known transmission chains have the potential to elucidate the impact of transmission dynamics on the viral evolutionary rate and its difference within and between hosts. Lin et al. (2015, *Journal of Virology*, 89/7: 3512–22) recently investigated the evolutionary history of hepatitis B virus in a transmission chain and postulated that the 'colonization–adaptation–transmission' model can explain the differential impact of transmission on synonymous and non-synonymous substitution rates. Here, we revisit this dataset using a full probabilistic Bayesian phylogenetic framework that adequately accounts for the non-independence of sequence data when estimating evolutionary parameters. Examination of the transmission chain data under a flexible coalescent prior reveals a general inconsistency between the estimated timings and clustering patterns and the known transmission history, highlighting the need to incorporate host transmission information in the analysis. Using an explicit genealogical transmission chain model, we find strong support for a transmission-associated decrease of the overall evolutionary rate. However, in contrast to the initially reported larger transmission effect on non-synonymous substitution rate, we find a similar decrease in both non-synonymous and synonymous substitution rates that cannot be adequately explained by the colonization-adaptation-transmission model. An alternative explanation may involve a transmission/establishment advantage of hepatitis B virus variants that have accumulated fewer within-host substitutions, perhaps by spending more time in the covalently closed circular DNA state between each round of viral replication. More generally, this study illustrates that ignoring phylogenetic relationships can lead to misleading evolutionary estimates.

Key words: hepatitis B virus; substitution rate; transmission chain; statistical phylogenetics; BEAST.

Importance

Pathogen strains carrying mutations that confer a selective advantage are expected to become fixed within a population. For the hepatitis B virus (HBV), however, mutations that confer adaptation within a single host are generally not passed on to the next host, which suggests that there are conflicting selective constraints for within-host survival versus transmission.

Determining the strategy that HBV uses to reconcile these different fitness requirements is of general interest to evolutionary virologists as it can help explain why rates of molecular change are dependent on the evolutionary time scale on which they are measured. By using methods that appropriately take the shared ancestry into account we find that, contrary to what was previously reported, strains that are less adapted to their current host, perhaps by undergoing fewer rounds of viral replication, are preferentially transmitted.

1. Introduction

Reliable estimates of evolutionary rates are critical for testing hypotheses in molecular epidemiology, involving, for example, the timing of viral emergence (Faria et al. 2016) or associating past demographic changes with ecological factors (Bryant et al. 2007; Markov et al. 2012; Trovão et al. 2015; Gill et al. 2016; Al-Qahtani et al. 2017; Cuypers et al. 2017). Various statistical descriptions of sequence divergence through time, collectively referred to as 'clock models', have been proposed to estimate the rate of nucleotide substitution (see Ho and Duchêne 2014 for a comprehensive overview). These models require calibration information in order to root phylogenies in time. If a significant amount of evolutionary change has accumulated over the sampling time span, the difference in sample isolation dates can be used to calibrate the tick rate of the molecular clock (Rambaut 2000; Drummond et al. 2003b; Biek et al. 2015). Over the years, it has become clear that the estimated rate of evolution depends on the evolutionary timescale (Ho et al. 2011, 2005; Aiewsakun and Katzourakis 2016), implying that the rate of change estimated from recently collected samples may not be applicable to deeper evolutionary time scales. This so-called time-dependency of evolutionary rates is also central to the debate about the evolutionary origins of hepatitis B virus (HBV), where discrepancies in evolutionary rate estimates complicate distinguishing between competing scenarios. Specifically, when the molecular clock is calibrated using recent sampling dates, the estimated evolutionary rates appear too high to be compatible with the 'out of Africa' hypothesis (Fares and Holmes 2002), whereas this hypothesis finds support when calibration relies on human-HBV co-divergence information (Paraskevis et al. 2013).

In addition to model misspecification (Ho et al. 2011), unaccounted for evolutionary constraints and saturation problems (Holmes 2003; Wertheim and Kosakovsky Pond 2011; Duchêne et al. 2014; Bielejec et al. 2014), a mismatch between the within- and between-host evolutionary rate (Zhou and Holmes 2007) can also contribute to the time dependency of molecular clock estimates. Indeed, transient within-host polymorphisms lead to inflated within-host substitution rate estimates because it is typically assumed that all polymorphisms are fixed (Duffy et al. 2008). When sampling occurs close to transmission, such effects may not impose a strong rate mismatch because there is little opportunity to accumulate many host-specific substitutions. HBV patients, however, usually only develop liver disease decades after infection, and, consequently, most patients are only sampled late in infection. Furthermore, a rate mismatch will confound dating efforts in phylogenies calibrated by dated tips for which the sampling time period covers only a small fraction of the total evolutionary history. This may be particularly true for HBV, which is assumed to have co-evolved with modern humans for the past ~40,000 years (Paraskevis et al. 2013), while samples are available only for the last few decades (with the notable exception of a HBV genotype C2 sequence recovered from a 16th century Korean mummy, Kahila Bar-Gal et al. 2012).

Known transmission chains provide unique opportunities to investigate what causes may underlie a difference between within- and among-host evolutionary rates. This has recently attracted considerable attention for the human immunodeficiency virus-1 (HIV-1), for which three hypotheses have been put forward to explain a lower evolutionary rate between hosts (Lythgoe and Fraser 2012). According to the 'store-and-retrieve' hypothesis, variants that avoid participating in the within-host evolutionary arms race better preserve the key phenotypic characteristics that enable to efficiently establish new infections and are hence more likely to do so. As 'stored' variants undergo fewer rounds of replication than actively replicating lineages they accumulate fewer mutations and the between-host rate will be lower than the within-host rate (Pybus and Rambaut 2009; Lythgoe and Fraser 2012). Alternatively, the rate discrepancy can also be explained by frequent transmission before the virus is subjected to strong immune-driven selection ('stage-specific selection') (Maljkovic Berry et al. 2007) or by the frequent reversion of adaptive mutations ('adapt-and-revert') (Herbeck et al. 2011). Importantly, these hypotheses make different predictions about the extent to which transmission affects the synonymous (μ_S) and non-synonymous (μ_N) contribution to the evolutionary rate. Specifically, when more slowly evolving variants have a transmission advantage, transmission will equally affect μ_S and μ_N whereas it is expected that μ_N contributes less to the overall rate at the between-host than at the within-host level under the two alternative scenarios. By investigating the signal in sequence data, different studies were able to attribute the HIV-1 rate slow-down to the reservoir dynamics (Alizon and Fraser 2013; Vrancken et al. 2014).

Lin et al. (2015) analysed the evolutionary dynamics in a known HBV transmission chain and found evidence for an evolutionary rate decline with increasing transmission, supporting the idea that different evolutionary processes act at the within- and between-host level (Zhou and Holmes 2007). They also observed a higher impact of transmission on the non-synonymous component of the evolutionary rate. To explain this, the authors consider the different stages of the virus life cycle, in which transmission is followed by colonization and then adaptation to the new host, and which pose conflicting phenotypic demands to the virus. They propose that variants that are well-adapted to the previous host may be less fit in the new host and therefore be outcompeted by variants that do not carry the specific adaptive changes. Although conceptually identical to the 'adapt-and-revert' hypothesis, Lin et al. (2015) coin this the 'colonization-adaptation-transmission' (CAT) model to emphasise the need for a continual switching between a set of phenotypes during the course of infection.

Here, we revisit this dataset and embed the sequence data in a full probabilistic Bayesian phylogenetic framework (BEAST, Drummond et al. 2012) to evaluate the robustness of their inferences against methods that properly take the correlation structure due to shared ancestry into account, which is ignored by pairwise comparisons to calculate evolutionary rates as used in the original study. Specifically, pairwise measurements may count substitutions on particular branches multiple times, which can impact rate estimates with unpredictable biases (Drummond et al. 2003a).

2. Materials and Methods

We follow a workflow previously applied to a known HIV transmission chain (Vrancken et al. 2014) and first examine the congruence between the timings of coalescent events, the

Table 1. Overview of the data and phylogenetic model combinations.

	Transmission history	Among lineage rate comparison			Transmission-effect on μ_S and μ_N	
	Unaware ('unconstrained')	Trunk[a]	Trunk[a]	Branch[b]	Within-host	Between-host
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Data	All	All	All	All	D1–3, GD2–3	Earliest sample
Coalescent model	Skygrid	Skygrid	Constant[c]	Constant[c]	Constant	Constant[c]
Clock model	ucld	ME	ME	ME	Strict	Strict
HPM[d]	–	–	–	–	Coalescent, clock	–

ucld, uncorrelated relaxed clock model where branch rates are drawn from a discretised lognormal distribution.
[a]The fixed effect that allows for a possibly different substitution rate on a subset of branches is specified on the transmitted lineages ('trunk').
[b]The fixed effect that allows for a possibly different substitution rate on a subset of branches is specified on the branches that accommodate one or more transmission events ('branch').
[c]This refers to the within-patient coalescent process under the transmission model, which is shared among all individuals of the transmission chain.
[d]The parameters for which information is pooled through hierarchical prior specification.

clustering patterns and the known sequence of timed transmission events in a standard time-measured phylogenetic analysis (further referred to as the 'unconstrained' analysis). Next, we explicitly test for a transmission-associated evolutionary rate decline using a mixed-effects (ME) molecular clock model while imposing transmission compatibility in our genealogical inference. Finally, we estimate the contribution of synonymous and non-synonymous substitutions to the overall evolutionary rate at the within- and between-host levels.

2.1 Datasets

We analyse the known HBV transmission chain data that was recently described by Lin et al. (2015). This chain covers seven vertical transmission events: from a grandmother (GM) to her five children (three daughters, D1, D2, D3, and two sons, S1 and S2), and from two of the daughters (D2 and D3) to the granddaughters (GD2 and GD3). We used the available complete genomes generated by cloning and Sanger sequencing of HBV variants from all members of this chain. We refer to Supplementary Fig. S1 for more details on the family relations, the times of infection (which equal the birth dates) and the sampling times and to Lin et al. (2015) for a more complete description of the data. Sequences were aligned using Muscle (Edgar 2004) and manually edited in AliView (Larsson 2014).

All sequence data were used in the 'unconstrained' analyses and in the analyses in which among lineage rate variation is captured using the ME relaxed clock model. For the direct comparison of the contribution of μ_S and μ_N to the within- and between-host evolutionary rates, we focused on subsets of the data. Specifically, only patients for which samples of multiple time points are available (all daughters and granddaughters) were included in the within-host evolutionary rate estimation, while only the sample closest to the transmission event of each patient was included in the between-host rate estimation (Table 1).

2.2 Bayesian phylogenetic inference

We parameterised the substitution process using an HKY substitution model (Hasegawa et al. 1985) and modelled among-site rate variation using a discretised Γ-distribution (Yang 1993) in the 'unconstrained' analyses and when estimating the within- and between-host rates. The skygrid model (Gill et al. 2013) was used as a flexible tree prior for all 'unconstrained' analyses. For the among host evolutionary rate estimations and for contrasting evolutionary rates among lineages, we accommodate the transmission history as in Vrancken et al. (2014). Briefly, the transmission model combines temporal constraints (the most recent common ancestor or MRCA of all donor and recipient lineages is constrained to predate the time of transition) with host transition compatibility requirements (correct host jumps are required upon transmission) to ensure the viral genealogy is fully compatible with the known transmission history. As detailed previously, this model does not make any assumptions about the transmission bottleneck. Because of the sparse within-host sampling, we resorted to the coalescent model with the fewest parameters (constant population size) to describe the demographic process within-hosts. We used different molecular clock models in our analyses. Evolutionary rates were estimated using an uncorrelated relaxed clock with rates drawn from a lognormal distribution (Drummond et al. 2006) in the 'unconstrained' analyses. The ME model was used to test for a transmission-associated rate slow-down for a specified set of internal branches. This clock model combines random effects on branch-specific evolutionary rates modelled according to an uncorrelated relaxed clock process with estimable fixed effects on pre-specified subsets of branches (akin to fixed local molecular clock modelling) (Vrancken et al. 2014). Finally, the simplest strict clock model was used for the independent estimations of the between- and within-host evolutionary rates. To obtain informative estimates for the sparser within-host data, we allow for sharing of information about population sizes and strict clock parameters in these analyses by specifying hierarchical phylogenetic models (HPM) (Suchard et al., 2003).

We follow Lin et al. (2015) in considering only the non-overlapping regions to obtain estimates of synonymous (μ_S) and non-synonymous (μ_N) substitution rates. Briefly, this involves summarising estimates of the absolute number of synonymous (S) and non-synonymous (N) substitutions (Minin and Suchard 2008; O'Brien et al. 2009; Lemey et al. 2012) as rates by dividing the total S and N counts by the alignment length and the total tree length in time units (Vrancken et al. 2014).

Posterior estimates under the full probabilistic model were obtained using Markov Chain Monte Carlo sampling as implemented in BEAST (Drummond et al. 2012), which was used in conjunction with BEAGLE (Ayres et al. 2012). Convergence and mixing properties of the chains were inspected using Tracer v1.6 (http://tree.bio.ed.ac.uk). Maximum clade credibility trees were summarised using the TreeAnnotator tool in BEAST and

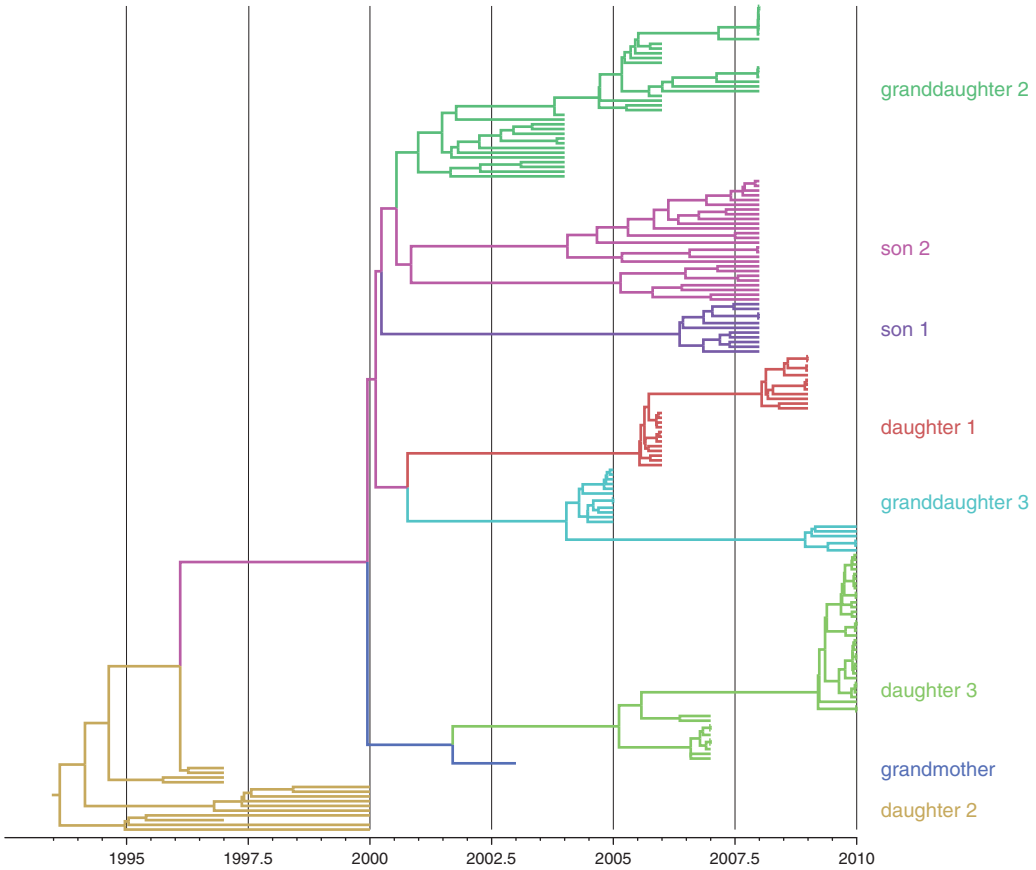

Figure 1. HBV phylogeny estimated from the transmission chain data under the 'unconstrained' model. We refer to Table 1 for details on the model setup. To facilitate identifying incompatibilities with the transmission history, we colored the tips and internal branches using a discrete asymmetric trait analysis with the patients as discrete states (Lemey et al. 2009; Edwards et al. 2011). The patient-color links are as indicated next to the tree.

visualised in FigTree v1.4.2 (http://tree.bio.ed.ac.uk). An overview of the dataset and phylogenetic model combinations is presented in Table 1.

3. Results

3.1 Conflicts between the genealogy and the transmission history

To investigate the level of compatibility between the viral genealogy and the known transmission history, we superimpose the known transmission history onto the reconstructed viral genealogy and summarise transmission compatibility of each transmission event as described previously by Vrancken et al. (2014). The estimated t_{MRCA} of the viral genealogy, 1992 (95% HPD: 1990–1994) in the 'unconstrained' analysis (model 1, Table 1) shows that calibrating the molecular clock with samples collected from the more recent part of the history—the sampling time span covers only the last 13 years of the virus history that goes back at least 66 years—results in a misleadingly high estimate of the evolutionary rate (1.86×10^{-3} substitutions/site/year, 95% HPD: $1.47–2.19 \times 10^{-3}$). As a consequence all donor–recipient lineages coalesce too early in time to be compatible with the known transmission history (Fig. 1). This is further aggravated by a host transition conflict involving GM, D2, and GD2. The basal clustering of the sequences from D2, which are the oldest sequences sampled from the transmission chain, makes a scenario of transmission from GM to D2, and subsequently from D2 to GD2, implausible (Fig. 1).

The general incompatibility between the viral genealogy and the transmission history underscores the need to incorporate external information into the analysis in order to obtain realistic parameter estimates (model 2, Table 1). Interestingly, imposing compatibility between the viral genealogy and the known chain of infections through time requires the hypothesis that D2 and D3 were infected by two variants. In both cases, the variant that was eventually transmitted to GD2 and GD3, respectively, was either not sampled or did not lead to any progeny (Fig. 2).

3.2 Evolution slows at transmission

To test for a transmission-associated rate slow-down, we applied a ME molecular clock model that allows for a different rate along the lineages that are transmitted from the index case all the way down into the different recipients compared to the rate on the branches that represent within patient evolution (model 3, Table 1 and Supplementary Fig. S2A). This allows testing whether the lineages that are being transmitted are the ones that accumulate fewer mutations as compared to the 'dead-end' within-host branches. Although the taxa specific to each family member always formed monophyletic clades when enforcing transmission history compatibility, we now explicitly enforced this to make sure that the fixed effect was always

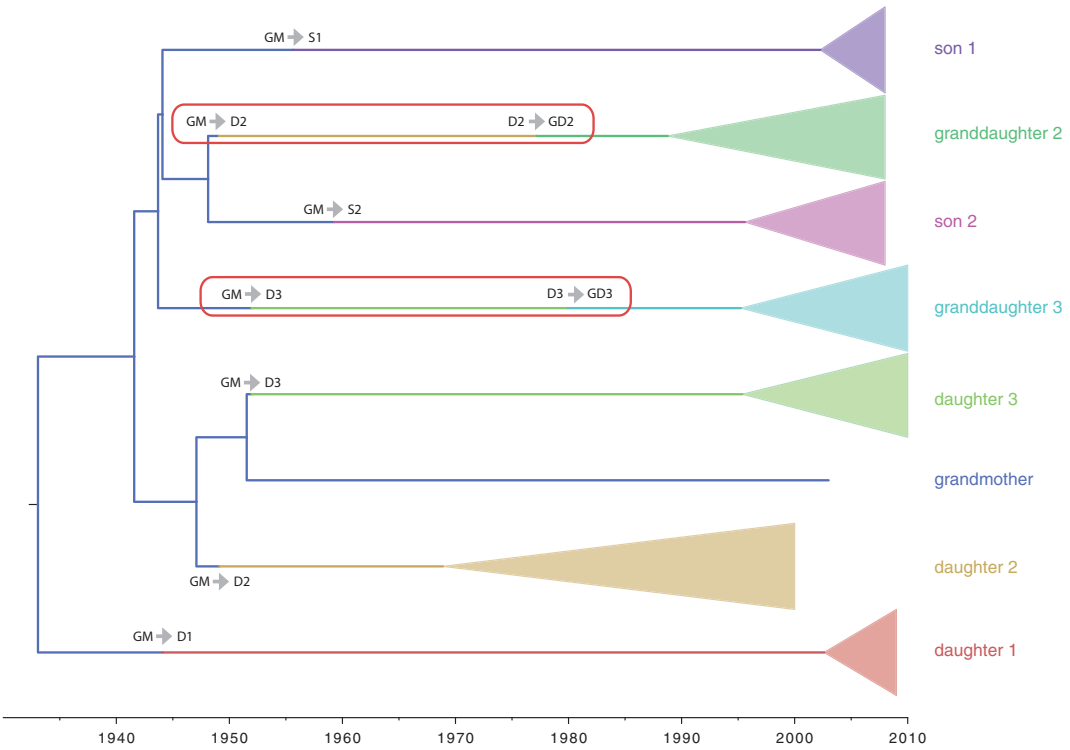

Figure 2. HBV phylogeny estimated from the transmission chain data while imposing compatibility with the transmission history. We refer to Table 1 for details on the model setup. The same color representation of patients is used as in Fig. 1. Each transmission event on a branch is represented by a color change. Multiple host jumps over the same branch are highlighted by red boxes.

Table 2. Overview of ME model evolutionary rate analyses.

Demographic model	Fixed-effects specification		Within-host	ln BF[a]	Fold change[b]
	'Trunk'	'Branch'			
Transmission model	0.76 (0.33–1.25)	n.a.	8.91 (6.02–11.8)	>10.57	11.7
	n.a.	0.57 (0.24–0.99)	8.36 (5.82–11.1)	>8.26	14.7
Skygrid ('unconstrained')	8.1 (3.5–13.5)	n.a	19.1 (14.7–23.7)	5.56	2.4

The mean evolutionary rate and HPD intervals are expressed as the number of nucleotide substitutions (10^{-4}) per site per year. 'Trunk' refers to the ME model specification in which a possibly different substitution rate operates on the transmitted lineages (this corresponds to model 3 in Table 1). 'Branch' refers to the ME model specification in which a possibly different substitution rate operates on the branches that accommodate one or more transmission events (this corresponds to model 4 in Table 1).

[a]The natural logarithm of the Bayes factor estimate in favour of a slower evolutionary rate on the transmitted lineages or the branches that accommodate one or more transmission events compared to the within-host rate.

[b]The fold change of the mean evolutionary rate estimate over the within-host lineages with respect to the mean evolutionary rate estimate over the between-host lineages.

associated with the branches that link the within-host populations. In agreement with the original study, we find a transmission-associated decrease in the evolutionary rate and recover strong support for a lower rate on the transmitted lineages, which we express as the natural logarithm of the Bayes factor estimate (ln BF, see Table 2). We also applied an ME model that specifies a rate effect on the branches that connect different hosts (model 4 in Table 1 and Supplementary Fig. S2B) and find strong support for a lower rate on branches that accommodate a transmission event (Table 2). The estimated rate on this subset of branches is also somewhat lower than that for the transmitted lineages (Table 2 and Supplementary Fig. S3). To test for the significance of this rate difference, we further considered an analysis combining fixed effects on the transmitted lineages and the transmission-accommodating branches. The ln BF in support of a lower substitution rate on the latter class of branches is 3.96, which is usually considered as strong evidence (Kass and Raftery 1995).

We also checked whether the well-supported non-zero rate difference could be an artefact of imposing the transmission model constraints by combining the 'trunk' ME model with the skygrid tree prior (model 2, Table 1) instead of the transmission model (model 3, Table 1). This shows that imposing the transmission model constraints decreases the substitution rate, but the ln Bayes factor support >5 for a rate difference (Table 2) reassures that the temporal and host transition constraints do not introduce the effect although they make the difference more pronounced.

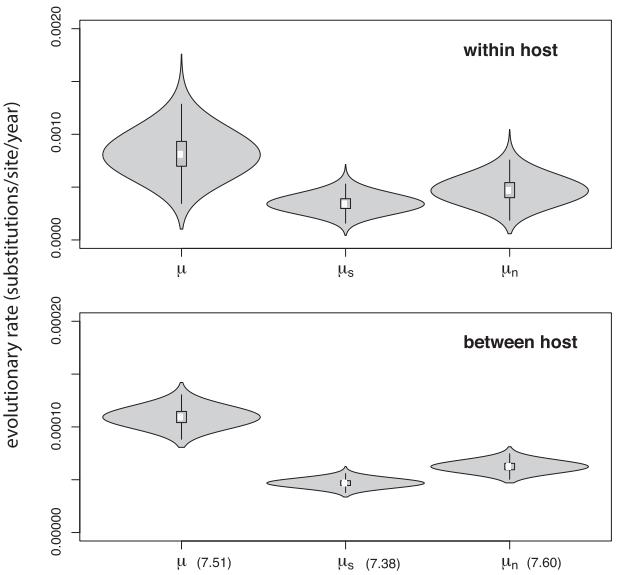

Figure 3. Substitution rate estimates within and between hosts. We refer to Table 1 for the model setup for the within- and between-host rate estimates. Numbers between parentheses indicate the fold decrease of the mean relative to the within-host mean rate estimate.

3.3 Synonymous and non-synonymous substitution rates are similarly affected by transmission

To gain insight into the mechanisms underlying the rate difference, we used a robust counting approach (Minin and Suchard 2008; O'Brien et al. 2009; Lemey et al. 2012) to estimate the synonymous (μ_S) and non-synonymous (μ_N) contribution to the evolutionary rate in the non-overlapping reading frames. Figure 3 shows the posterior probability densities for the overall rate μ, μ_S and μ_N for both the within- and between-host analysis. The scale of the Y-axis illustrates the rate difference between both biological scales, with within-host rate estimates being an order of magnitude higher than between-host estimates (10^{-3} compared to 10^{-4}). Importantly, Fig. 3 also shows that μ_S and μ_N are affected similarly by transmission, which stands in contrast with the reported larger effect of transmission on μ_N by Lin et al. (2015).

We suspect that this discrepancy can be attributed to the use of different estimation procedures. Due to shared ancestry, pairwise comparisons will count substitutions over particular branches multiple times, potentially biasing the distance estimates that involve such branches (Lemey et al. 2009). To explore whether such biases are present, we coloured all branches of the maximum clade credibility tree inferred under the transmission model by their associated difference in expected number of N and S mutations (Fig. 4). This shows that a select number of branches have a pronounced positive difference between N and S counts. The most notable of these is the branch leading to the GM because substitutions along this branch are expected to be counted in five (GM→D1, D2, D3, S1, S2) of the seven (the other two are D2→GD2 and D3→GD3) distance estimates between clades separated by one transmission event, and it is this category that accounts for the previously detected effect (see Fig. 3 in Lin et al. 2015). The substitutions along this branch are also not counted for within-host estimates because only a single sequence is available for the GM. We note that the difference in expected N and S counts over this branch is highly similar to the estimated difference in an unconstrained analysis (Supplementary Fig. S4).

4. Discussion

In this study, we revisited the impact of transmission on the HBV substitution rate in a known transmission chain previously examined by Lin et al. (2015). Here, we analysed the genetic data using models that appropriately take into account shared ancestry. This is important because, as shown by our results, ignoring the evolutionary relationships can introduce unpredictable biases and lead to misleading results.

Using a standard demographic model as tree prior, we found that the divergence time estimates of donor–recipient lineages are in conflict with the timed series of known transmission events. This is likely explained by peculiarities of the HBV biology and the fact that the HBV patients were chronically infected for decades before the virus population was sampled (range, 25–80 years). The latter is important because, while the HBeAg prevents the immune system from mounting a response during early chronic infection, this virus antigen is usually not expressed at later stages (Harrison et al. 2011). The preCore 1986 G→A substitution is a hallmark for the $HBeAg^-$ status, and was found in 76.7 per cent of the sampled lineages (Supplementary Table S2). Therefore, the evolutionary divergence that accumulated over the sampling period to some extent reflects fast $HBeAg^-$ status associated evolution and the too recent divergence time estimates can at least be partly explained by the impact of the immune system on the HBV evolutionary rate.

In addition to anomalous timings, the 'unconstrained' analysis also revealed a discordant clustering pattern (Fig. 1). As was noted by Lin et al. (2015), this can be a consequence of the quick reversion of mutations that were adaptive in the previous host environment, such that infection is initiated by similar variants in all individuals. The decades of evolution between infection and the first sampling event in immunologically similar hosts too may have resulted in homoplasies that confound the inferred clustering patterns, making them incompatible with the transmission history.

Next, we explicitly tested for a transmission-associated evolutionary rate decline using a ME molecular clock model while imposing coalescent compatibility under the transmission model. The ME clock model allows for among-branch rate variation in the same way as a standard relaxed clock model. Two observations reassure us that this provides sufficient flexibility to adequately capture the variation in within-host rates that is most likely linked to the HBeAg status. First, the estimated substitution rate over the transmitted lineages and branches that accommodate a transmission event, which likely represent the $HBeAg^+$ periods of infection, is consistently at the lower end of the rate estimate spectrum (Table 2). Second, the 95 per cent HPD interval of the between-host rate largely overlaps with those estimated from data sets restricted to $HBeAg^+$ carrier sequences (Harrison et al. 2011).

Like Lin et al. (2015), we find a transmission-associated decrease in the evolutionary rate, and recover strong Bayes factor support for a lower rate on the transmitted lineages (Table 2). Because the switch from $HBeAg^+$ to $HBeAg^-$ usually occurs between the ages of 15 and 35 years (Merican et al. 2000), which is more or less the donor age at transmission in this chain (Supplementary Table S1), it is very likely that the donor was $HBeAg^+$ for some transmission events. Since HBV evolves slower during the $HBeAg^+$ stage of infection, and our sampling partly reflects faster $HBeAg^-$ evolution, it could be that the rate difference detected by the ME clock model between the

Figure 4. HBV phylogeny estimated from the transmission chain data using the non-overlapping parts of the open reading frames. The tree was inferred using only the first sample per patient and the transmission model (model 6, Table 1). The width and color of branches accord to the difference in the expected number of non-synonymous and synonymous mutations (E[N]-E[S]) they accommodate. For clarity, only values >3 are shown. The difference on the branch leading to the GM is highlighted with a red circle.

transmitted lineages or branches that accommodate a transmission event, on the one hand, and the branches that represent within-host evolution, on the other hand, is linked to the HBeAg status and not to impact transmission. However, we find strong support (ln BF = 3.96) for a lower substitution rate on the branches that accommodate a transmission event than on the transmitted lineages (Table 2 and Supplementary Fig. S3). This indicates that in addition to a potential effect of $HBeAg^+$ evolution on the 'trunk' or transmitted lineages, we find support for a specific effect of slower evolution associated with transmission.

Because HBV usually is transmitted by $HBeAg^+$ carriers, it has been suggested that the slower evolution during the $HBeAg^+$ phase appropriately approximates the long-term substitution rate of the virus and could be used for studying events over longer periods of time (Simmonds 2001). However, they are still an order of a magnitude higher than the rate estimates obtained by calibrating the molecular clock using human-virus co-divergence data (2.2×10^{-6}, 95%HPD: $1.5–3.0 \times 10^{-6}$, Paraskevis et al. 2013), suggesting that saturation and long-term purifying selection (Holmes 2003; Wertheim and Kosakovsky Pond 2011; Duchêne et al. 2014; Bielejec et al., 2014) may play a role at longer evolutionary time scales. This time-dependency of substitution rates has permeated the field of molecular evolution and advocates the use of internal calibrations to obtain more appropriate divergence times over longer time scales (Worobey et al. 2008, 2010; Paraskevis et al. 2013; Duggan et al. 2016).

The difference in evolutionary rates was also apparent in the direct comparison of the within- and between-host estimates (Fig. 3), where we evaluated the effect of transmission on μ_S and μ_N. The interpretation of the impact of transmission on μ_S and μ_N measurements critically depends on the assumption that synonymous changes are less affected by selection than non-synonymous changes. Synonymous changes are not per se selectively neutral, for example, due to their impact on the secondary RNA and DNA structure, and the compact nature of the HBV genome may further aggravate this. However, it has been shown for HIV-1, which also has a compact genome, that the sites that affect the secondary RNA structure are under strong purifying selection, which indicates that this class of sites contributes little to the substitution rate (Sanjuán and Bordería 2011). In addition, available data for single-stranded viruses indicate that synonymous changes have markedly lower selection coefficients than amino acid changing mutations (Cuevas et al. 2012).

In contrast to Lin et al. (2015), we found that μ_S and μ_N experience a similar decline, and we indicate how ignoring the tree structure can bias pairwise divergence estimates. Under the CAT model put forward by Lin et al. (2015), it is expected that the non-synonymous changes that have accumulated within a host have reverted by the time of sampling in the new host, and are thus not accounted for in the between host substitution rate measurements. Because synonymous changes are expected to be less affected by selection, the CAT model predicts that transmission should have a more pronounced impact on the non-synonymous component of the substitution rate. This conflicts with our results, which instead are in line with the recently proposed hypothesis by Lythgoe et al. (2017); that is, HBV variants mitigate the trade-off between more efficient replication within single hosts and decreased transmissibility between hosts by spending a long time in the covalently closed circular DNA (cccDNA) state, which can persist for prolonged periods of time

(Yang and Kao 2014), between rounds of replication. In this way, the viral generation time increases and, hence, the opportunity to accumulate deleterious adaptive mutations decreases (Lythgoe et al. 2017). To account for the differences in evolutionary rates within and between hosts, we further propose that viruses that have undergone fewer rounds of replication, and therefore had less opportunity to accumulate mutations, are more likely to be transmitted (similar to what has been proposed for HIV and HCV). Interestingly, a recent survey of 37 mother-to-child-transmission pairs found that the majority of new infections (∼75%) was initiated by the predominant maternal variant (Yang et al. 2017). If new infections are indeed usually initiated by those variants that have spent more time in the cccDNA state between rounds of replication, this implies that faster replicating variants usually are also evolutionary dead-ends within hosts. More research, for example, in the form of long-term longitudinal follow-up studies similar to those performed for hepatitis C virus (Raghwani et al. 2016), is needed to further examine whether the observed within-host dynamics match the evolutionary patterns that are predicted by this hypothesis.

More generally, this study demonstrates why evolutionary hypothesis testing should preferentially be based on probabilistic models that take the correlation structure into account that is induced by shared ancestry and that model sequence evolution as realistically as possible.

Acknowledgements

This study was supported by the Bijzonder Onderzoeksfonds KU Leuven (BOF) No. OT/14/115 and the Fonds voor Wetenschappelijk Onderzoek Vlaanderen (FWO) (G066215N). The VIROGENESIS project receives funding from the European Unions Horizon 2020 research and innovation program under grant agreement No 634650. The research leading to these results has received funding from the European Research Council under the European Community's Horizon 2020 Programme under ERC Grant agreement no. 725422. M.A.S. was partially supported by National Science Foundation grant DMS 1264153 and National Institutes of Health grants R01 AI107034 and R01 AI117011.

Data availability

The previously generated virus genetic data can be found in GenBank, accession numbers KP406161–KP406335.

Supplementary data

Supplementary data are available at *Virus Evolution* online.

Conflict of interest: None declared.

References

Aiewsakun, P., and Katzourakis, A. (2016) 'Time-dependent rate phenomenon in viruses', *Journal of Virology*, 90/16: 7184–95

Al-Qahtani, A. A. et al. (2017) 'The epidemic dynamics of hepatitis c virus subtypes 4a and 4d in Saudi Arabia', *Science Report*, 7: 44947

Alizon, S., and Fraser, C. (2013) 'Within-host and between-host evolutionary rates across the HIV-1 genome', *Retrovirology*, 10/1: 49

Ayres, D. L. et al. (2012) 'BEAGLE: an application programming interface and high-performance computing library for statistical phylogenetics', *Systematic Biology*, 61/1: 170–3

Biek, R. et al. (2015) 'Measurably evolving pathogens in the genomic era', *Trends in Ecology and Evolution*, 30/6: 306–13

Bielejec, F. et al. (2014) 'πBUSS: a parallel BEAST/BEAGLE utility for sequence simulation under complex evolutionary scenarios', *BMC Bioinformatics*, 15: 133

Bryant, J. E., Holmes, E. C., and Barrett, A. D. T. (2007) 'Out of Africa: a molecular perspective on the introduction of yellow fever virus into the Americas', *PLoS Pathogens*, 3/5: e75

Cuevas, J. M., Domingo-Calap, P., and Sanjuán, R. (2012) 'The fitness effects of synonymous mutations in dna and rna viruses', *Molecular Biology and Evolution*, 29/1: 17–20

Cuypers, L. et al. (2017) 'Implications of hepatitis c virus subtype 1a migration patterns for virus genetic sequencing policies in italy', *BMC Evolutionary Biology*, 17/1: 70

Drummond, A., Pybus, O. G., and Rambaut, A. (2003a) 'Inference of viral evolutionary rates from molecular sequences', *Advances in Parasitology*, 54: 331–58

Drummond, A. J. et al. (2003b) 'Measurably evolving populations', *Trends in Ecology and Evolution*, 18/9: 481–8

—— et al. (2006) 'Relaxed phylogenetics and dating with confidence', *PLoS Biology*, 4/5: e88

—— et al. (2012) 'Bayesian phylogenetics with BEAUti and the BEAST 1.7', *Molecular Biology and Evolution*,

Duchêne, S., Holmes, E. C., and Ho, S. Y. W. (2014) 'Analyses of evolutionary dynamics in viruses are hindered by a time-dependent bias in rate estimates', *Proceedings Biological Sciences*, 281/1786:

Duffy, S., Shackelton, L. A., and Holmes, E. C. (2008) 'Rates of evolutionary change in viruses: patterns and determinants', *Nature Review Genetics*, 9/4: 267–76

Duggan, A. T. et al. (2016) '17(th) century variola virus reveals the recent history of smallpox', *Current Biology*, 26/24: 3407–12

Edgar, R. C. (2004) 'MUSCLE: a multiple sequence alignment method with reduced time and space complexity', *BMC Bioinformatics*, 5: 113

Edwards, C. J. et al. (2011) 'Ancient hybridization and an Irish origin for the modern polar bear matriline', *Curr Biol*, 21/15: 1251–8

Fares, M. A., and Holmes, E. C. (2002) 'A revised evolutionary history of hepatitis B virus (HBV)', *Journal of Molecular Evolution*, 54/6: 807–14

Faria, N. R. et al. (2016) 'Zika virus in the Americas: early epidemiological and genetic findings', *Science*, 352/6283: 345–9

Gill, M. S. et al. (2013) 'Improving Bayesian population dynamics inference: a coalescent-based model for multiple loci', *Molecular Biology and Evolution*, 30/3: 713–24

—— et al. (2016) 'Understanding past population dynamics: Bayesian coalescent-based modeling with covariates', *Systematic Biology*, 65/6: 1041–56

Harrison, A. et al. (2011) 'Genomic analysis of hepatitis B virus reveals antigen state and genotype as sources of evolutionary rate variation', *Viruses*, 3/2: 83–101

Hasegawa, M., Kishino, H., and Yano, T. (1985) 'Dating of the human-ape splitting by a molecular clock of mitochondrial DNA', *Journal of Molecular Evolution*, 22/2: 160–74

Herbeck, J. T. et al. (2011) 'Demographic processes affect hiv-1 evolution in primary infection before the onset of selective processes', *Journal of Virology*, 85/15: 7523–34

Ho, S. Y. W., and Duchêne, S. (2014) 'Molecular-clock methods for estimating evolutionary rates and timescales', *Molecular Ecology*, 23/24: 5947–65

—— et al. (2005) 'Time dependency of molecular rate estimates and systematic overestimation of recent divergence times', *Molecular Biology and Evolution*, 22/7: 1561–8

—— et al. (2011) 'Time-dependent rates of molecular evolution', *Molecular Ecology*, 20/15: 3087–101

Holmes, E. C. (2003) 'Molecular clocks and the puzzle of RNA virus origins', *Journal of Virology*, 77/7: 3893–7

Kahila Bar-Gal, G. et al. (2012) 'Tracing hepatitis B virus to the 16th century in a Korean mummy', *Hepatology*, 56/5: 1671–80

Kass, R. E., and Raftery, A. E. (1995) 'Bayes factors', *Journal of the American Statistical Association*, 90/430: 773–95

Larsson, A. (2014) 'Aliview: a fast and lightweight alignment viewer and editor for large datasets', *Bioinformatics*, 30/22: 3276–8

Lemey, P., Salemi, M., and Vandamme, A.-M., eds (2009). *The Phylogenetic Handbook*, chapter 14, pp. 419–451. Cambridge: Cambridge University Press.

—— et al. (2012) 'A counting renaissance: combining stochastic mapping and empirical bayes to quickly detect amino acid sites under positive selection', *Bioinformatics*, 28/24: 3248–56

Lin, Y.-Y. et al. (2015) 'New insights into the evolutionary rate of hepatitis B virus at different biological scales', *Journal of Virology*, 89/7: 3512–22

Lythgoe, K. A., and Fraser, C. (2012) 'New insights into the evolutionary rate of HIV-1 at the within-host and epidemiological levels', *Proceedings Biological Sciences*, 279/1741: 3367–75

—— et al. (2017) 'Short-sighted virus evolution and a germline hypothesis for chronic viral infections', *Trends Microbiology*, 25/5: 336–48

Maljkovic Berry, I. et al. (2007) 'Unequal evolutionary rates in the human immunodeficiency virus type 1 (HIV-1) pandemic: the evolutionary rate of HIV-1 slows down when the epidemic rate increases', *Journal of Virology*, 81/19: 10625–35

Markov, P. V. et al. (2012) 'Colonial history and contemporary transmission shape the genetic diversity of hepatitis C virus genotype 2 in Amsterdam', *Journal of Virology*, 86/14: 7677–87

Merican, I. et al. (2000) 'Chronic hepatitis b virus infection in asian countries', *Journal of Gastroenterology and Hepatology*, 15/12: 1356–61

Minin, V. N., and Suchard, M. A. (2008) 'Fast, accurate and simulation-free stochastic mapping', *Philosophical Transactions of the Royal Society B: Biological Sciences*, 363/1512: 3985–95

O'Brien, J. D., Minin, V. N., and Suchard, M. A. (2009) 'Learning to count: robust estimates for labeled distances between molecular sequences', *Molecular Biology and Evolution*, 26/4: 801–14

Paraskevis, D. et al. (2013) 'Dating the origin and dispersal of hepatitis B virus infection in humans and primates', *Hepatology*, 57/3: 908–16

Pybus, O. G., and Rambaut, A. (2009) 'Evolutionary analysis of the dynamics of viral infectious disease', *Nature Review Genetics*, 10/8: 540–50

Raghwani, J. et al. (2016) 'Exceptional heterogeneity in viral evolutionary dynamics characterises chronic hepatitis c virus infection', *PLoS Pathogens*, 12/9: e1005894

Rambaut, A. (2000) 'Estimating the rate of molecular evolution: incorporating non-contemporaneous sequences into maximum likelihood phylogenies', *Bioinformatics*, 16/4: 395–9

Sanjuán, R., and Bordería, A. V. (2011) 'Interplay between RNA structure and protein evolution in hiv-1', *Molecular Biology and Evolution*, 28/4: 1333–8

Simmonds, P. (2001) 'The origin and evolution of hepatitis viruses in humans', *Journal of General Virology*, 82/Pt 4: 693–712

Suchard, M. A. et al. (2003) 'Hierarchical phylogenetic models for analyzing multipartite sequence data', *Systematic Biology*, 52/5: 649–64

Trovão, N. S. et al. (2015) 'Host ecology determines the dispersal patterns of a plant virus', *Virus Evolution*, 1/1:

Vrancken, B. et al. (2014) 'The genealogical population dynamics of HIV-1 in a large transmission chain: bridging within and among host evolutionary rates', *PLOS Computational Biology*, 10/4: e1003505

Wertheim, J. O., and Kosakovsky Pond, S. L. (2011) 'Purifying selection can obscure the ancient age of viral lineages', *Molecular Biology and Evolution*, 28/12: 3355–65

Worobey, M. et al. (2008) 'Direct evidence of extensive diversity of HIV-1 in Kinshasa by 1960', *Nature*, 455/7213: 661–4

—— et al. (2010) 'Island biogeography reveals the deep history of SIV', *Science*, 329/5998: 1487

Yang, Z. (1993) 'Maximum likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites', *Molecular Biology and Evolution*, 10: 1396–401

Yang, H.-C., and Kao, J.-H. (2014) 'Persistence of hepatitis B virus covalently closed circular DNA in hepatocytes: molecular mechanisms and clinical significance', *Emerging Microbes & Infections*, 3/9: e64

Yang, G. et al. (2017) 'Quasispecies characteristics in mother-to-child transmission of hepatitis b virus by next-generation sequencing', *Journal of Infection*, 75/1: 48–58

Zhou, Y., and Holmes, E. C. (2007) 'Bayesian estimates of the evolutionary rate and age of hepatitis B virus', *Journal of Molecular Evolution*, 65/2: 197–205