Virus Evolution, 2016, 2(2): vew023

doi: 10.1093/ve/vew023
Research article

Identifying predictors of time-inhomogeneous viral evolutionary processes

Filip Bielejec[1,*], Guy Baele[1], Allen G. Rodrigo[2,†], Marc A. Suchard[3,4], and Philippe Lemey[1,‡]

[1]Department of Microbiology and Immunology, Rega Institute, KU Leuven, Leuven, Belgium, [2]Research School of Biology, Australian National University, Canberra, ACT, Australia, [3]Department of Biomathematics and 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: filip.bielejec@rega.kuleuven.be

†http://orcid.org/0000-0002-8327-7317

‡http://orcid.org/0000-0003-2826-5353

Abstract

Various factors determine the rate at which mutations are generated and fixed in viral genomes. Viral evolutionary rates may vary over the course of a single persistent infection and can reflect changes in replication rates and selective dynamics. Dedicated statistical inference approaches are required to understand how the complex interplay of these processes shapes the genetic diversity and divergence in viral populations. Although evolutionary models accommodating a high degree of complexity can now be formalized, adequately informing these models by potentially sparse data, and assessing the association of the resulting estimates with external predictors, remains a major challenge. In this article, we present a novel Bayesian evolutionary inference method, which integrates multiple potential predictors and tests their association with variation in the absolute rates of synonymous and non-synonymous substitutions along the evolutionary history. We consider clinical and virological measures as predictors, but also changes in population size trajectories that are simultaneously inferred using coalescent modelling. We demonstrate the potential of our method in an application to within-host HIV-1 sequence data sampled throughout the infection of multiple patients. While analyses of individual patient populations lack statistical power, we detect significant evidence for an abrupt drop in non-synonymous rates in late stage infection and a more gradual increase in synonymous rates over the course of infection in a joint analysis across all patients. The former is predicted by the immune relaxation hypothesis while the latter may be in line with increasing replicative fitness during the asymptomatic stage.

Key words: Bayesian phylogenetics; evolutionary rate; pathogen; virus evolution; generalized linear models; codon substitution models; epoch models.

1. Introduction

Evolutionary and population genetic processes in rapidly evolving viruses can be highly dynamic even on human observable timescales. Quantifying these dynamics and testing how ecology and host immune responses shape them is a major objective of phylodynamic analyses. Coalescent approaches, for example, allow inferring changes in population size through time from genealogies (Drummond et al. 2005), and recent modelling advances have demonstrated the ability to incorporate ecological complexities,

such as vector dynamics and spatial structure, that can affect the shape of viral genealogies (Rasmussen, Boni and Koelle 2014). To assess changes in the tempo or mode of sequence evolution, most modelling efforts focus on variation among branches, either by accommodating specific prior hypotheses (e.g. clade-specific evolutionary rates modelled according to different hosts, see Worobey, Han and Rambaut 2014) or by adopting flexible approaches that allow identifying the changes. Both approaches are available to detect or test variation in molecular clock rates (Drummond and Suchard 2010; Worobey, Han and Rambaut 2014) and variation in non-synonymous/synonymous substitution rate ratios (dN/dS, see Yang 1998).

Identifying changes in dN/dS allows uncovering heterogeneity in selective pressure. In order to examine how such patterns vary in HIV-1 evolution over the course of a single infection, Goode, Guindon and Rodrigo (2008) proposed an approach that models evolutionary shifts in the codon substitution process that cut across all lineages at a specific time point in a time-calibrated genealogy. We have recently presented a Bayesian implementation of such a time-inhomogeneous model that generalizes 'epoch' specifications to any discrete substitution process in an unknown evolutionary history, including, for example, phylogeographic processes (Bielejec et al. 2014). The epoch model allows specifying an arbitrary sequence of continuous-time Markov chain (CTMC) models through time while appropriately accommodating phylogenetic uncertainty. Likelihood computations under an epoch scenario require matrix convolution which adds to the high computational burden associated with fitting high-state space models to relatively large data sets. To considerably speed up these calculations, we have implemented massively parallel likelihood computations on multi-core devices such as graphics processing units.

In order to test prior hypotheses about the viral divergence stabilization observed in the AIDS stage of HIV-1 infection, we previously applied epoch modelling to extensively sampled sequence data over the time course of infection in several individuals (Bielejec et al. 2014). Specifically, we conditioned on the progression time for each patient, which represents the time at which the CD4+ T-cell counts drop below 200 cells/μl (Williamson 2003), to partition the evolutionary history into 2 intervals (the asymptomatic stage and the AIDS stage). By estimating dN/dS in independent codon substitution models associated with both epochs, we found a generally lower selective pressure on the viral population after progression time (in the AIDS stage). We interpreted this pattern as support for the 'immune relaxation hypothesis', which attributes the lower selective pressure to a damaged immune stage in the AIDS stage (Williamson 2003).

Although we were able to address this hypothesis using an epoch model with prior-determined transition times, disease progression may not necessarily follow a discrete two-step process. In fact, CD4+ T-cell counts on which the progression times are based generally decrease continuously throughout infection history, and inversely correlated with this, the viral load increases over time. If the evolutionary process is correlated with disease progression, it may be hypothesized that evolutionary parameters will follow the evolution of clinical parameters. In general, this calls for a framework to test potential predictors of the evolutionary process through time. For independent realizations of the evolutionary process (e.g. HIV-1 evolution in different patients), we have previously demonstrated the ability to estimate the support and contribution of explanatory variables to evolutionary parameters in an integrated Bayesian framework (Edo-Matas et al. 2011; Streicker et al. 2012).

Although estimating dN/dS has proven useful in an epoch context (Bielejec et al. 2014), this only allows us to detect changes in selection dynamics. Intrahost viral evolution in persistent infections may also be influenced by factors that impact the underlying mutation rate or generation rate, such as variation in replication rates which was shown to be associated with differences in HIV-1 disease progression among patients (Lemey et al. 2007), or HIV-1 storage in latently infected cells which may explain differences in evolutionary rate within and between patients (Vrancken et al. 2014). Disentangling these factors from selection forces requires a separate estimate of synonymous and non-synonymous substitution rates. In a fixed-tree molecular clock framework this may be achieved by codon modelling approaches aimed at estimating absolute rates of synonymous and non-synonymous substitution (Seo, Kishino and Thorne 2004). In Bayesian inference approaches that accommodate phylogenetic uncertainty, the computational burden associated with codon models has restricted their application to relatively large datasets. This has motivated the development of different post-hoc procedures (Lemey et al. 2007) and nucleotide-based counting proxies (Lemey et al. 2012). While these approximations aim to quantify synonymous and non-synonymous substitution rates, they do not treat these quantities as model parameters and are therefore less suitable to extend to formal hypothesis testing procedures. With the current ability to considerably speed up likelihood calculations (Suchard and Rambaut 2009), massively parallel computation is now stimulating further codon model development in the Bayesian framework.

Here, we use an implementation of the Muse & Gaut (MG94) codon substitution model (Muse and Gaut 1994) in BEAST (Drummond et al. 2012) to estimate synonymous and non-synonymous substitution rates. We adopt the approach in an epoch setting and further extend it to accommodate potential predictors of the substitution process using generalized linear modelling. We employ this framework to identify the support and contribution of various predictors of synonymous and non-synonymous substitution rates in intrahost HIV evolution. We do not find sufficient signal in separate analyses of eight different patients, but when jointly estimating predictor support and effect size, we find that progression time remains the best explanatory variable for non-synonymous rate variation despite the fact that we also consider CD4+ T-cell count and viral load measurements.

2. Methods

2.1 Codon substitution modelling

Our approach builds on the standard MG94 codon substitution model (Muse and Gaut 1994), which is parameterized in terms of a synonymous (α) and non-synonymous (β) substitution rate by defining a CTMC infinitesimal rate matrix $\mathbf{Q} = \{q_{ij}\}$ with the following off-diagonal entries:

$$q_{ij} = \begin{cases} \alpha\kappa\pi_j & i \to j \text{ is a one} - \text{nucleotide synonymous transition from codon } i \text{ to } j. \\ \alpha\pi_j & i \to j \text{ is a one} - \text{nucleotide synonymous transversion from codon } i \text{ to } j. \\ \beta\kappa\pi_j & i \to j \text{ is a one} - \text{nucleotide non} - \text{synonymous transition from codon } i \text{ to } j. \\ \beta\pi_j & i \to j \text{ is a one} - \text{nucleotide non} - \text{synonymous transversion from codon } i \text{ to } j. \\ 0 & \text{otherwise}, \end{cases} \quad (1)$$

where κ is the transition–transversion rate and $\pi_{j\ j=1,\ldots,61}$ denotes the frequency of the target codon.

We fit this codon model in a Bayesian framework and use Markov chain Monte Carlo (MCMC) integration to obtain a sample from the posterior distribution of model parameters (Drummond et al. 2012). Because BEAST infers trees in time-units (e.g. using tip or node calibrations), substitution model matrices are generally normalized (to one substitution per site per unit time) and scaled by an estimable overall evolutionary rate parameter into the time units of the tree. In order to estimate absolute rates of synonymous and non-synonymous substitution rates, we perform a normalization that gives rise to $\alpha + \beta$ expected substitutions per site per year.

2.2 Epoch and general linear modelling

In order to identify predictors of changes in α and β through time, we first embed the codon substitution process into an epoch setting (Bielejec et al. 2014). This setting posits that the substitution process is homogeneous within an epoch of time, but may change across epochs. To construct these epochs, we set, without loss of generality, the time of the most recent sequence sample from each patient to 0 and let the remaining $M-1$ sampling-times define transition times $T_1, \ldots, T_{M-1}$, where we assume that time flows backwards, that is from the tips towards the MRCA. Together with $T_M = \infty$, the ordered times $T_0 < T_1 < \cdots < T_M$ define the boundaries of M epochs, such that epoch m begins at time T_{m-1} and ends at time T_m. Note that the last epoch extends to ∞, but with the evolutionary history only extending to the MRCA.

For each of the M epochs, we associate a conditionally independent, infinitesimal rate matrix $\mathbf{Q}_m = \mathbf{Q}_m(\alpha_m, \beta_m)$ with $m = 1, \ldots, M$ that depends on epoch-specific synonymous α_m and non-synonymous β_m rates. Then, for any two arbitrary times $t_u \leq t_v$ under this time-inhomogeneous process, we can compute the matrix of finite-time transition probabilities

$$\mathbf{P}(t_u, t_v) = \prod_{m=1}^{M} \exp\{\mathbf{Q}_m[\min(t_v, T_m) - \max(t_u, T_{m-1})]^+\}, \quad (2)$$

where $[\cdot]^+ = \max(\cdot, 0)$. While the computational cost of Equation (2) appears high, it is important to note that when the time-interval $[t_u, t_v]$ does not intersect with a given epoch m',

$$[\min(t_v, T_{m'}) - \max(t_u, T_{m'-1})]^+ = 0, \text{ and } \exp\{\mathbf{Q}_{m'} \times 0\} = \mathbf{I}, \quad (3)$$

where $\mathbf{I}$ is the identity matrix and, as such, no matrix-exponentiation nor matrix-multiplication is required for that epoch. Second, we parameterize α_m and β_m as log-linear functions of potential predictors:

$$\log \alpha_m = \mu_\alpha + \mathbf{X}_m(\theta_\alpha \odot \delta_\alpha), \text{ and } \log \beta_m = \mu_\beta + \mathbf{X}_m(\theta_\beta \odot \delta_\beta), \quad (4)$$

where μ_k is the grand-mean rate (on the log-scale) for $k \in \{\alpha, \beta\}$, $\mathbf{X}_m = (X_{m1}, \ldots, X_{mP})$ are P predictor values for epoch m, $\theta_k = (\theta_{k1}, \ldots, \theta_{kP})^T$ quantify the contribution of each predictor to rate k, $\delta_k = (\delta_{k1}, \ldots, \delta_{kP})^T$ are binary indicator variables that model the in-/exclusion of the predictor and $\odot$ is the component-wise product operation. To ensure that μ_k quantifies the average rate across all epochs, we standardize the predictor values such that $(X_{1p}, \ldots, X_{Mp})$ has mean 0 and standard deviation 1 for all quantifiable entries across $p = 1, \ldots, P$. Never quantifiable values may arise, for example, in the last epoch when predictors are functions of measurements taken at the sampling-time that demarcates the end of an epoch. For such predictors, we fix $X_{Mp} = 0$ after standardization, so that $\log \alpha_M = \mu_\alpha$ and $\log \beta_M = \mu_\beta$.

To complete our prior specification, we assume that:

$$\begin{aligned} &(\mu_\alpha, \mu_\beta) \sim \text{MVN}(0, 1000 \times \mathbf{I}), \\ &\theta_k \sim \text{MVN}\left(0, \frac{1}{7} \times \mathbf{I}\right) \text{ for all } k, \text{ and} \\ &\delta_{kp} \sim \text{Bernoulli}(q) \text{ for all } k \text{ and } p, \end{aligned} \quad (5)$$

where $\text{MVN}(\cdot, \cdot)$ signifies a multivariate normal distribution with given mean and variance and p is a prior inclusion probability. These priors incorporate the belief that most predictors will only have a modest impact on the evolutionary rate, but we wish to remain uninformative about its overall size through the grand-means. We select q such that there exists a 50% prior probability that no predictor is included in the model. Augmenting the GLM parameterisation with binary indicator variables for the predictors and their associated prior specification allows for a Bayesian stochastic search variable selection (Lemey et al. 2009) procedure, which estimates posterior probabilities of inclusion or exclusion of a particular predictor in Equation (4) and allows us to readily compute Bayes factor support. In our analyses, we summarize mean and the 95% highest posterior density interval for the conditional effect size, which is the size of the effect (on log scale) conditional on the effect being included in the model ($\theta_{kp} | \delta_{kp} = 1$).

2.3 Sequence data and predictors

We re-analyse extensively serially sampled *env* C2V5 sequences from eight patients throughout their infection starting close to the time of seroconversion (Shankarappa et al. 1999). We do not consider one particular patient from the original study (patient 11) for which no data were available after progression time. The sequence data consist of the C2-V5 region of the HIV-1 *env* gene sequences collected in a longitudinal manner over a 6–13.7-year period, with a minimum of 5 and a maximum of 15 measurements per patient. In total, the data constitutes 1,300 sequences and 106 separate time points.

Most sampling times were associated with CD4+ T-cell counts and viral load (VL) (Supplementary Fig. 1), which we use as predictors in two different ways upon log transformation: either as the mean of the measurements at two boundaries of each epoch (*mean CD4* and *mean VL*) or as the difference between the two boundary values ($\Delta CD4$ and ΔVL). We also use the progression time as predictor, which is encoded as a binary indicator (in log space, with the same estimable effect size for each epoch prior to this time). In this sense, it may be more appropriate to refer to this as infection stage predictor. Furthermore, we consider several sampling characteristics as predictors such as the number of sequences sampled (the mean number of sequences sampled at both boundaries), the time since seroconversion (measured from the midpoint of the epoch) and the time length for each epoch (*epoch time*).

Finally, we design a way to include effective population size change (ΔN_e) over the epoch as a predictor for the sequence evolutionary parameters. For this purpose, we employ the recently developed Bayesian Skygrid model (Gill et al. 2013) and match the grid intervals to our epoch structure. We incorporate the difference in log N_e at the boundaries of the closed intervals as a predictor in our GLM design matrix. We opt to standardize predictors in our GLM approach and implement a dynamic

standardization mechanism for the predictors in the design matrix because the log N_e-based predictors are continuously updated during the MCMC. We note that the use of parameter estimates of the tree-generative model as predictors for the sequence evolutionary process informing the tree may be problematic. Initial explorations indeed indicated that this leads to an identifiability problem for the tree prior and sequence evolutionary parameters. To avoid this, we use an empirical set of trees, estimated by a standard nucleotide sequence analysis in BEAST, and average over this distribution while estimating population size and sequence evolutionary parameters.

Several observations are missing for the CD4 and VL variables, specifically for the datasets for patients 3, 7, 8 and 9. We used a feed-forward neural network with one hidden layer with ten nodes to impute those values. This graph of interconnected nodes (neurons) is capable of learning by adjusting the weights of the paths connecting its inputs to outputs (Bryson and Ho 1969). In the datasets for patients 3 and 9 several observations are missing for both predictors, and since the neural network is trained using backward propagation of errors, we recode all the values to fall between $(-1, 1)$ with missing values coded as 0, such that the weights for these inputs are also shrunk to 0 during back-propagation and effectively no learning is done on those branches.

2.4 Joint estimation

In addition to fitting the GLM-model to the data from each patient, we also aim to obtain a joint estimate of predictor support and effect size across in a hierarchical phylogenetic setting (Suchard et al. 2003). In this analysis, we specify independent tree parameters and non-parametric coalescent processes for each patient. For the sequence evolutionary process, we specify shared predictor indicators and coefficients in the GLM-parameterisations and shared codon equilibrium frequencies. In addition, we specify hierarchical prior distributions (Edo-Matas et al. 2011) to pool information for the κ parameters, the shape parameters of the gamma distributions for the among-site rate variation and the skygrid precision parameters. We further include patient-specific random-effects on the grand-means μ_α and μ_β to account for absolute substitution rate differences across patients.

3. Results

We apply our approach to HIV-1 *env* C2V3 sequences sampled throughout the infection of eight patients (Shankarappa et al. 1999) in order to estimate α and β and identify correlates for variation in these rates through time. We adopt an epoch structure that follows the sampling-times and consider clinical and virological parameters as well as sampling characteristics and changes in N_e as predictors for the dynamics of synonymous and non-synonymous substitution rates.

We summarize independent estimates of predictor inclusion probability ($\mathbb{E}[\delta_{kp}]$) for each predictor and each patient as a stacked barplot in Figure 1. In general, this demonstrates very little support for any predictor to consistently explain variation in either α or β. As an indication, the bar plot shows a vertical line at eight times the individual prior probability of predictor inclusion. For α, only progression time and time since seroconversion yield a sum of inclusion probabilities that is somewhat higher than this value, but for both, about half of this sum is contributed by a single patient. For β, both progression time and mean CD4 show some elevation in the probability sum, but not substantial enough to attribute importance to the estimates. The only apparent consistency is that the correlated measures of progression time and mean CD4—as the former is the time at which the CD4+ T-cell counts drop below 200 cells/μl (Williamson 2003)— are both elevated for β.

To increase the statistical power and inform our model simultaneously by all the patient-specific data, we set up a joint analysis with shared predictor indicators and effect sizes (cfr. Section 2). Figure 2 summarizes the posterior inclusion probabilities and the corresponding contribution of each predictor to the

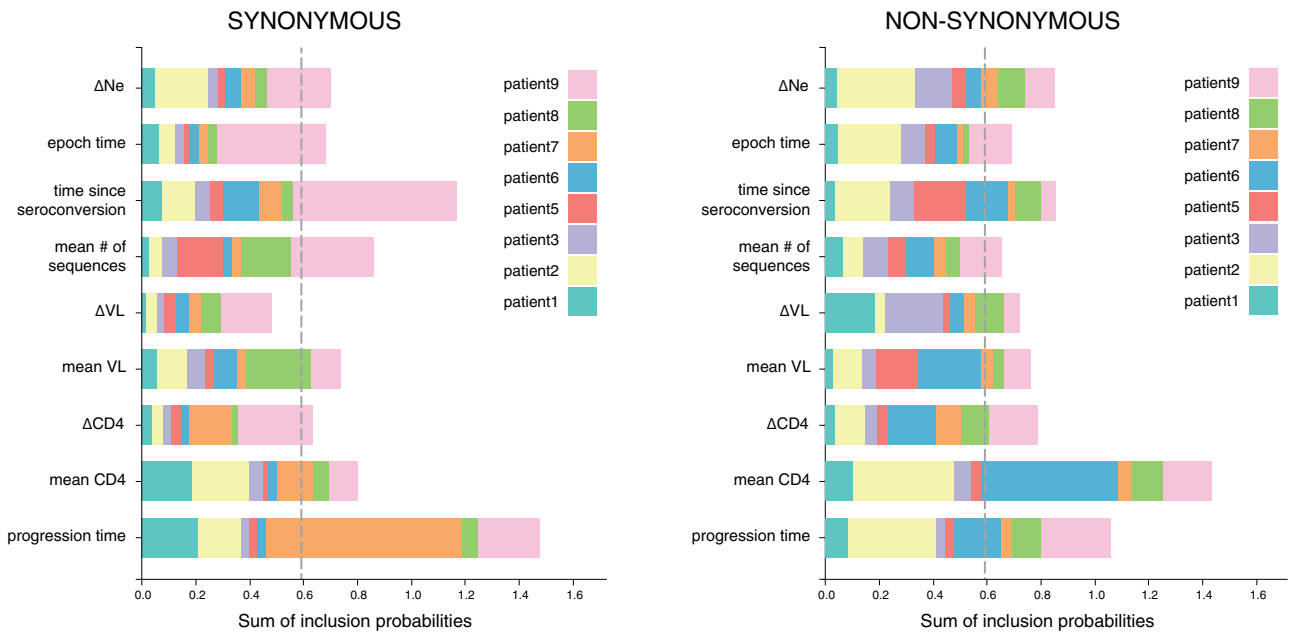

Figure 1. Predictor inclusion probabilities for each predictor and each patient. All predictor inclusion probabilities estimated separately from each patient-specific viral population are summarized into a stacked bar plot. The bar plot on the left and right represents the summed inclusion probabilities for predictors of α and β, respectively. The vertical dashed line represents eight times the prior inclusion probability used in each individual analysis.

dependent variable (the posterior mean and credible interval for that coefficient, conditioning on its inclusion in the GLM: that is, $(\theta_{kp}|\delta_{kp} = 1)$). We find strong support for the time since seroconversion to predict variation in α (posterior inclusion probability of 0.88, Bayes Factor value of 89.96), with a positive conditional effect size 0.15 (0.07, 0.22) indicating a general increase in α over the course of the infection. In addition, the epoch time also yields non-negligible support (posterior inclusion probability of 0.45, Bayes Factor value of 10.47), but with a negative effect size −0.08 (−0.14, −0.02), suggesting a negative correlation between α and epoch length. For β, we find strong support for progression time (posterior inclusion probability of 0.95, Bayes factor value of 232.2), with a positive effect size 0.15 (0.07, 0.22). Because the progression time predictor was encoded as a homogeneous estimable effect before this time point, it corresponds to higher β's before disease progression. We also note the moderate support for ΔN_e (posterior inclusion probability of 0.21, Bayes factor value of 3.4), with a positive effect size 0.1 (0.02, 0.19).

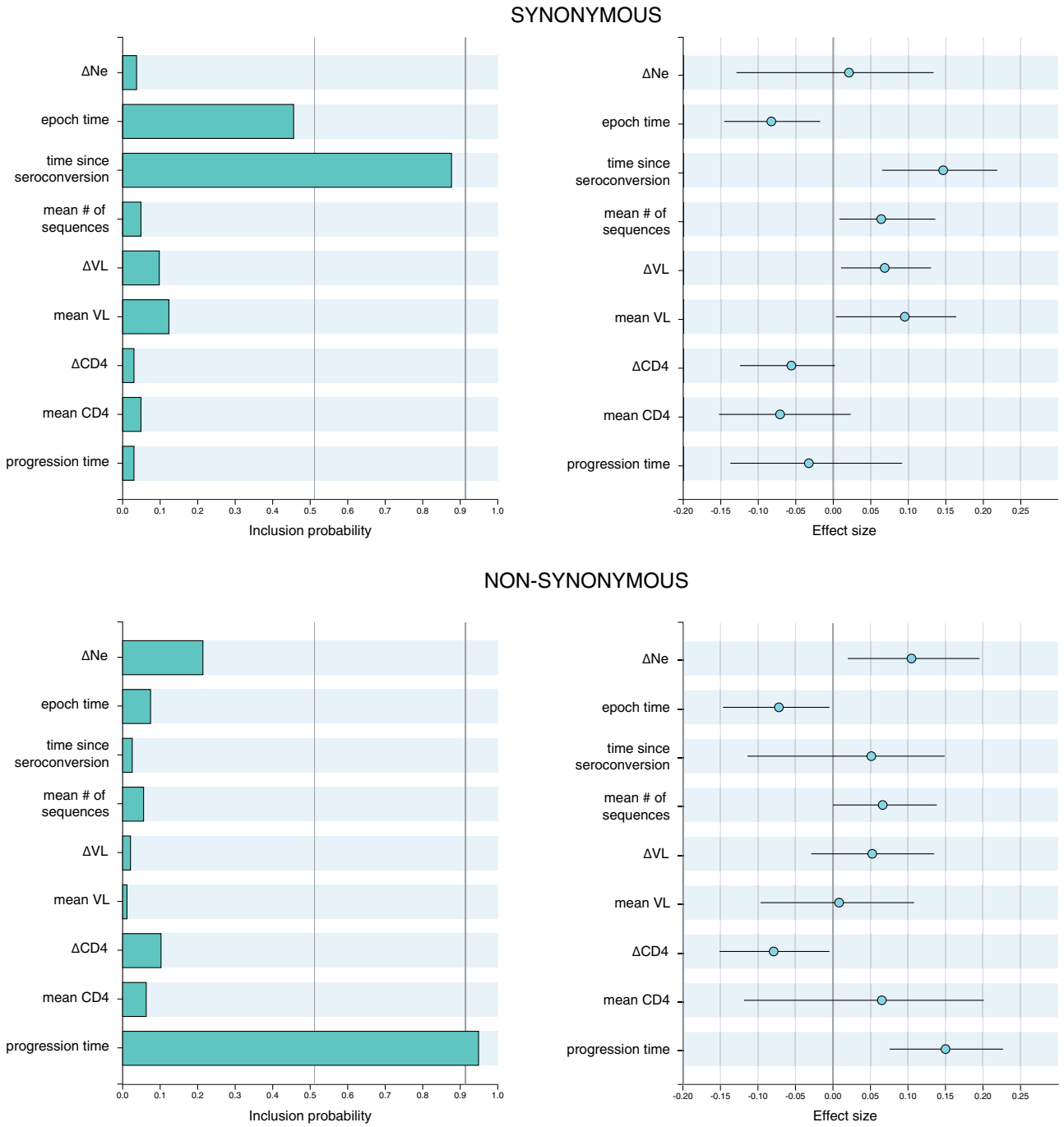

Figure 2. Joint inclusion probabilities and conditional effect sizes on the log scale for the predictors of α or β. The inclusion probabilities (plots on the left) and conditional effect sizes ($\theta|\delta = 1$, plots on the right) are shared across all patients. The upper and bottom plots represent the estimates for α or β, respectively. Thin and thick black vertical lines in the barplot present δ indicator expectations corresponding to Bayes factor values of 10 and 100, respectively, which following Kass and Raftery (1995) can be interpreted as 'substantial' and 'very strong' evidence.

4. Discussion

In this study, we present a novel approach to test associations between changes in substitution rates over time with changes in external predictors. The approach builds on previous developments that allow changing substitution processes through time (Goode, Guindon, and Rodrigo 2008; Bielejec et al. 2014), and also relates to ideas proposed by Rodrigo et al. (2008). To distinguish between factors that impact neutral rates of evolution and selective dynamics, we make use of a codon substitution model parameterized in terms of absolute synonymous and non-synonymous substitution rates. Our application to within-host HIV-1 evolution demonstrates the difficulty to extract signal for changes in these parameters from individual intrahost viral populations. However, joint inference over all patients leads to support for correlates for both the synonymous and non-synonymous substitution rates.

Since the initial observation of a slowdown or stabilization in divergence accumulation in the late phase of HIV-1 infection (Shankarappa et al. 1999), evolutionary theories have been put forward to explain these dynamics in relationship to disease progression. Williamson et al. (2005) formalized the two prevailing hypotheses as 'cellular exhaustion', or the reduced availability of target cells late in infection, and 'immune relaxation', which refers to reduced selective pressure because of deteriorating immune responses. These authors used measures of synonymous and non-synonymous evolution to test these hypotheses and found strong support for a cessation of non-synonymous divergence in line with immune relaxation. By applying a local codon model to a set of trees sampled from the posterior distribution of a Bayesian relaxed clock analysis in nucleotide space, Lemey et al. (2007) provided further evidence that non-synonymous divergence stabilises in some patients. The method we develop we here represents a coherent integration of codon substitution models in this Bayesian phylogenetic framework. We do not model branch-specific rate variation, but adopt epoch modelling to capture the variation in synonymous and non-synonymous rates through time. Our results also support a slowdown in non-synonymous substitution rates after progression time. Although progression time is defined as a threshold for the decline in CD4+ T-cell counts over time, we do not find a correlation with the underlying evolution of these counts. This may indicate that the immune collapse, and the associated drop in immune responses, is indeed a discrete event that occurs when the impairment rate of HIV exceeds the threshold value, as suggested by modelling studies of HIV-1 disease progression (Iwami et al. 2009; Huang, Takeuchi and Korobeinikov 2012).

Interestingly, our approach also finds support for a positive correlation between synonymous substitution rates and time since seroconversion, suggesting that neutral substitution rate is gradually increasing over time in HIV-1 infection. This is in line with an experimental study that found evidence for increasing HIV-1 replication efficiency over the course of the infection (Troyer et al. 2005). Increased replication rates may reduce the generation time in intrahost HIV-1 populations and lead to faster rates of neutral evolution. This has also been invoked as the explanation for the association between disease progression and synonymous substitution rates (Lemey et al. 2007). The posterior synonymous substitution rates through time suggest that the increase may not always be entirely consistent throughout infection (Supplementary Fig. 2), and in some patients this stabilizes towards the late stage of infection. This may explain why some have argued that cellular exhaustion may also sometimes be compatible with the divergence stabilization observed at the AIDS stage (Lee et al. 2008) . However, we note that synonymous rates stabilize in some patients, but do not really decrease, and the stabilization does not necessarily align with progression time (Supplementary Fig. 2). So, this may equally well reflect a natural cap on fitness increase or even a trade-off between immune escape and replication rate (Lemey et al. 2007).

It may prove interesting to apply our approach to other persistent infections in the future. Intrahost HCV populations for example exhibit strong heterogeneity in the rate of molecular evolution (Gray et al. 2011), perhaps due to variation in replication rate as suggested supported by mathematical models of HCV infection kinetics (Neumann et al. 1998). Together with population structure in the liver, this may induce complex chronic evolutionary patterns that are difficult to capture by simple statistics of viral genetic variation (Gray et al. 2012). Evolutionary rate variation and potential correlates have also been described in chronic HBV infections, but without making a distinction between synonymous and non-synonymous substitutions (Harrison et al. 2011). We note that our model is not limited to analysis of within-host pathogen genetic data and it readily extends to other problems and predictors. By implementing the method in the publicly available BEAST software (Drummond et al. 2012), it can be also connected to other models of sequence and traits evolution.

Future applications may benefit from a more coherent treatment of missing predictor data. We currently impute such predictors prior to our analysis, but it may be possible to integrate out the missing predictor values in our Bayesian inference framework. Furthermore, our model may be extended by introducing random effects for the evolutionary response variable in our GLM model. We have not pursued this in this study because the independent analyses of the patients already lacked statistical power. However, since the joint model was well informed by the data from all patients, random-effects may prove useful to explore in this context. Bayesian mixed-effects modelling has already been successfully applied in our framework (Edo-Matas et al. 2011; Streicker et al. 2012).

More generally, incorporating covariates may find various uses in phylodynamic inference. We consider population size estimates as covariates for the substitution process in this study, but it may also prove interesting to model N_e as a function of potential predictors, such as CD4 counts and viral load in within-host HIV-1 dynamics. This may also be relevant for models linking coalescent theory to compartmental models in epidemiology (Koelle and Rasmussen 2012; Volz, Koelle and Bedford 2013), which can also be applied to within-host HIV-1 dynamics.

An important area for future research with Bayesian codon model implementation lies in accommodating separate among-site and among-lineage variation in synonymous and non-synonymous substitution rates. Modelling separate variation in both quantities among sites has proven crucial in accurate detection of selection pressure (Kosakovsky Pond and Muse 2005), and this relates to a rich history in the development of site-specific selection detection methods in the maximum likelihood framework (Kosakovsky Pond, Poon and Frost 2009). Inferring branch-specific variation in synonymous and non-synonymous substitution rates can also deliver important evolutionary insights (Seo, Kishino and Thorne 2004), and this could be accommodated in our framework by connecting the codon substitution model to the uncorrelated relaxed clock models (Drummond et al. 2006). However, care will need to be

taken to keep the computational burden manageable as these models require an eigen-decomposition of the infinitesimal generator matrix for each branch when the tree is considered to be random, which is computationally challenging for high state-space models. Bayesian non-parametric priors specifically tailored for evolutionary problems might offer a solution because they allow identifying a limited number of rate classes, as has been demonstrated for nucleotide substitution rate variation among lineages (Huelsenbeck and Nielsen 1999). Finally, adequately modelling among-site and among-lineage variation in synonymous and non-synonymous substitution rates may also improve divergence dating in relatively deep viral phylogenies because nucleotide-based substitution models fail to account for complex patterns of spatial and temporal variability in selective pressures (Wertheim and Kosakovsky Pond 2011).

In conclusion, our novel approach to identify correlates synonymous and non-synonymous substitution rates confirms and refines previous hypotheses about intrahost HIV-1 evolutionary dynamics and provides the basis for promising extensions in Bayesian codon substitution modelling.

Supplementary data

Supplementary data are available at *Virus Evolution* online.

Funding

The research leading to these results has received funding from the European Union Seventh Framework Programme [FP7/2007-2013] under Grant Agreement No 278433-PREDEMICS and ERC Grant Agreement No 260864, from the National Institutes of Health (R01 AI107034, R01 HG006139 and LM011827) and the National Science Foundation (IIS 1251151 and DMS 1264153). The VIROGENESIS project receives funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 634650. F.B. is supported by a post-doctoral mandate from the Research Fund KU Leuven.

Conflict of interest: None declared.

References

Bielejec, F., et al. (2014) 'Inferring Heterogeneous Evolutionary Processes Through Time: From Sequence Substitution to Phylogeography', *Systematic Biology*, 63: 493–504.

Bryson, A. E. and Ho, Y.-C. (1969) *Applied Optimal Control: Optimization, Estimation, and Control*. Waltham: Blaisdell.

Drummond, A. and Suchard, M. (2010) 'Bayesian Random Local Clocks, or One Rate to Rule Them All', *BMC Biology*, 8: 114.

Drummond, A. J., et al. (2012) 'Bayesian Phylogenetics with BEAUti and the BEAST 1.7', *Molecular Biology and Evolution*, 29: 1969–73.

——, et al. (2005) 'Bayesian Coalescent Inference of Past Population Dynamics From Molecular Sequences', *Molecular Biology And Evolution*, 22: 1185–92.

——, et al. (2006) 'Relaxed Phylogenetics and Dating with Confidence', *PLoS Biology*, 4: e88.

Edo-Matas, D., et al. (2011) 'Impact of CCR5delta32 Host Genetic Background and Disease Progression on HIV-1 Intrahost Evolutionary Processes: Efficient Hypothesis Testing Through Hierarchical Phylogenetic Models', *Molecular Biology and Evolution*, 28: 1605–16.

Gill, M. S., et al. (2013) 'Improving Bayesian Population Dynamics Inference: A Coalescent-Based Model for Multiple Loci', *Molecular Biology and Evolution*, 30: 713–24.

Goode, M., Guindon, S., and Rodrigo, A. (2008) 'Modelling the Evolution of Protein Coding Sequences Sampled from Measurably Evolving Populations', *Genome Informatics*, 21: 150–64.

Gray, R. R., et al. (2011) 'The Mode and Tempo of Hepatitis c Virus Evolution Within and Among Hosts', *BMC Evolutionary Biology*, 11: 131.

——, et al. (2012) 'A New Evolutionary Model for Hepatitis c Virus Chronic Infection', *PLoS Pathogene*, 8: e1002656.

Harrison, A., et al. (2011) 'Genomic Analysis of Hepatitis b Virus Reveals Antigen State and Genotype as Sources of Evolutionary Rate Variation', *Viruses*, 3: 83–101.

Huang, G., Takeuchi, Y., and Korobeinikov, A. (2012) 'HIV Evolution and Progression of the Infection to Aids', *Journal of Theorteical Biology*, 307: 149–59.

Huelsenbeck, J. P., and Nielsen, R. (1999) 'Variation in the Pattern of Nucleotide Substitution Across Sites', *Journal of Molecular Evolution*, 48: 86–93.

Iwami, S., et al. (2009) 'Immune Impairment Thresholds in HIV Infection', *Immunology Letters*, 123: 149–54.

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

Koelle, K., and Rasmussen, D. A. (2012) 'Rates of Coalescence for Common Epidemiological Models at Equilibrium', *Journal of the Royal Society Interface*, 9: 997–1007.

Kosakovsky Pond, S. L., Poon, A. F. Y. S., and Frost D. W., (2009) *The Phylogenetic Handbook. A Practical Approach to Phylogenetic Analysis and Hypothesis Testing, chapter Estimating selection pressures on alignments of coding sequences*. Number 14. Cambridge: Cambridge University Press.

——, ——, and ——, and Muse, S. V. (2005) 'Site-to-Site Variation of Synonymous Substitution Rates', *Molecular Biology and Evolution*, 22: 2375–85.

Lee, H. Y., et al. (2008) 'Dynamic Correlation Between Intrahost HIV-1 Quasispecies Evolution and Disease Progression', *PLoS Computational Biology*, 4: e1000240.

Lemey, P., et al. (2007) 'Synonymous Substitution Rates Predict HIV Disease Progression as a Result of Underlying Replication Dynamics', *PLoS Computational Biology*, 3: e29.

——, et al. (2009) 'Bayesian Phylogeography Finds Its Roots', *PLoS Computational Biology*, 5: e1000520.

——, et al. (2012) 'A Counting Renaissance: Combining Stochastic Mapping and Empirical Bayes to Quickly Detect Amino Acid Sites Under Positive Selection', *Bioinformatics*, 28: 3248–56.

Muse, S. V. and Gaut, B. S. (1994) 'A Likelihood Approach for Comparing Synonymous and Nonsynonymous Nucleotide Substitution Rates, with Application to the Chloroplast Genome', *Molecular Biology and Evolution*, 11: 715–24.

Neumann, A. U., et al. (1998) 'Hepatitis c Viral Dynamics In Vivo and the Antiviral Efficacy of Interferon-Alpha Therapy', *Science*, 282: 103–7.

Rasmussen, D. A., Boni, M. F., and Koelle, K. (2014) 'Reconciling Phylodynamics with Epidemiology: The Case of Dengue Virus in Southern Vietnam', *Molecular Biology and Evolution*, 31: 258–71.

Rodrigo, A., et al. (2008) 'The Perils of Plenty: What Are We Going to Do With All These Genes?', *Philosophical Transactions of the Royal Society B: Biological Sciences*, 363: 3893–902.

Seo, T.-K., Kishino, H., and Thorne, J. L. (2004) 'Estimating Absolute Rates of Synonymous and Nonsynonymous

Nucleotide Substitution in Order to Characterize Natural Selection and Date Species Divergences', *Molecular Biology and Evolution*, 21: 1201–13.

Shankarappa, R., et al. (1999) 'Consistent Viral Evolutionary Changes Associated with the Progression of Human Immunodeficiency Virus Type 1 Infection', *Journal of Virology*, 73: 10489–502.

Streicker, D. G., et al. (2012) 'Rates of Viral Evolution Are Linked to Host Geography in Bat Rabies', *PLoS Pathogene*, 8: e1002720.

Suchard, M. A. and Rambaut, A. (2009) 'Many-Core Algorithms for Statistical Phylogenetics', *Bioinformatics*, 25: 1370–6.

——— and ———, et al. (2003) 'Hierarchical Phylogenetic Models for Analyzing Multipartite Sequence Data', *Systematic Biology*, 52: 649–64.

Troyer, R. M., et al. (2005) 'Changes in Human Immunodeficiency Virus Type 1 Fitness and Genetic Diversity During Disease Progression', *Journal of Virology*, 79: 9006–18.

Volz, E. M., Koelle, K., and Bedford, T. (2013) 'Viral Phylodynamics', *PLoS Computational Biology*, 9: e1002947.

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: 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: 3355–65.

Williamson, S. (2003) 'Adaptation in the Env Gene of HIV-1 and Evolutionary Theories of Disease Progression', *Molecular Biology and Evolution*, 20: 1318–25.

———, et al. (2005) 'A Statistical Characterization of Consistent Patterns of Human Immunodeficiency Virus Evolution Within Infected Patients', *Molecular Biology and Evolution*, 22: 456–68.

Worobey, M., Han, G. Z., and Rambaut, A. (2014) 'A Synchronized Global Sweep of the Internal Genes of Modern Avian Influenza Virus', *Nature*, 508: 254–7.

Yang, Z. (1998) 'Likelihood Ratio Tests for Detecting Positive Selection and Application to Primate Lysozyme Evolution', *Molecular Biology and Evolution*, 15: 568–73.