Bayesian Coalescent Inference of HCV and HIV-1B

Konrad Gjertsson[1], Nassim Versbraegen[2]

[1]Université Libre de Bruxelles, 1050 Brussels, Belgium
kgjertss@ulb.ac.be
[2]Vrije Universiteit Brussel, 1050 Brussels, Belgium
nversbra@vub.ac.be

Abstract

This article aims to reconstruct past population dynamics through coalescent theory and Bayesian inference. The methods used to achieve this goal are the Bayesian Skyline Plot and the Extended Bayesian Skyline Plot. Practical use cases in the field of epidemiology are discussed, and the techniques are applied to Egyptian gene sequences of Hepatitis C (HCV) and global samples of the human immunodeficiency virus (HIV). The examples in this study have been carefully selected inter alia due to the substantial amount of different, independent data, which allows for validation of the models. Our findings confirm that proper application of the methods leads to results corroborating real-world scenarios.

Keywords Coalescent theory, Bayesian Inference, Markov chain Monte Carlo, Skyline Plot, HCV, HIV-1B

Introduction

A number of undertakings have consistently fascinated mankind, one of which is the reconstruction of the past. In many cases this aim remains unattainable. However, for some applications, techniques have been developed to recover information which would otherwise have been lost to the ages. One such technique is Bayesian Coalescent Inference (Drummond et al., 2005). This technique allows for a reconstruction of past population dynamics, while employing current data. Uncovering a species' past can be beneficial in many cases and Bayesian Coalescent Inference is especially useful when no demographic record exists for certain organisms by which one could validate various hypotheses. Bayesian Coalescent Inference incorporates coalescent theory (Kingman, 1982b,a), and has had multiple significant updates since its introduction in 1982 (Oliver G. Pybus, 2000; Strimmer and Pybus, 2001; Drummond et al., 2005; Heled and Drummond, 2008). Coalescent theory is a relatively recent model as a consequence of its entanglement with sequencing techniques; inferences are based on genetic data. The model is stochastic and aims to relate genetic diversity in a given set of sequences to the demographic history of the population of which the sequences stem from (Harding, 1996). Although coalescent theory had been rigorously developed, it remained untouched for the use of inferring past population dynamics for 12 years after its introduction (Griffiths and Tavare, 1994; Donnelly and Tavare, 1995). The gist of the model's method consists of generating a hypothesis which incorporates genetic drift[1] and posits the coancestor (i.e. the most recent common ancestor), comparatively to phylogenetic relations, and the point in time at which divergence occurred (the moment of a split is called a coalescent event). One of coalescent theory's advantages is that it employs individual samples, unlike preceding methods, which required data about the entire population (in contrast with the more convenient sample of the population). Furthermore, estimation of the model can be attained through highly efficient algorithms (such as the Markov chain Monte Carlo method, as shown in e.g. the Bayesian Skyline Plot (BSP) (Drummond et al., 2005)) (Donnelly and Tavare, 1995). Through this research, we aspire to replicate and extend the endavour from Drummond et al's 2005 paper that introduced the Bayesian Skyline Plot. Additionally, this paper aims to employ the latest development in the field; the Extended Bayesian Skyline Plot (EBSP) (Heled and Drummond, 2008). This modus operandi allows for a comparison between both techniques.

HCV

An application of coalescent theory that is interesting due to its high practical value is the testing of hypotheses in epidemiology (Joy et al., 2003). Thus, as in Drummond et al. (2005); Heled and Drummond (2008), the two Bayesian inference methods have been applied to Egyptian samples of the Hepatitis C virus (HCV). Specifically, HCV type 4 and HCV type 1, subtype g have been included. HCV has been the predominant cause for chronic liver disease in Egypt since 1980 (Strickland, 2006). In 2011, 11 826 360 individuals (14% of the nations total population) were infected in the country (Lavanchy, 2011). The seroprevalence is about 10-20 times higher than in the United States (Ray et al.,

[1]Genetic variance between sequences (i.e. differences between nucleotide sequences) can be examined in combination with knowledge about typical allele distribution dynamics (e.g. exclusively genetic drift would lead to genetic uniformity) to infer which evolutionary phenomena accounts for the observed variance.

2000), and much higher than in neighboring countries. In addition, these samples show a number of desirable traits that make them remarkably suitable for coalescent analysis (Drummond et al., 2005):

- The samples are geographically diverse.
- There are no clear sub-population divisions.
- They contain adequate phylogenetic information.
- An independent estimate of nucleotide-substitution rate exists.

HIV

Additionally, our research includes an application of the method on HIV. Employed samples stem from the Western world and sub-Saharan Africa. The West is of particular interest due to the peculiar circumstances of the initial dissemination in the region, as will be elaborated upon in the discussion section. We deemed sub-Saharan Africa as a relevant source of sequences since current scientific understanding considers it to be the location of the initial infections.

HIV is a lentivirus, which is a part of the retroviruses (Desport, 2010). This entails that HIV causes DNA to be generated, based on it's own RNA strands (using it's reverse transcriptase molecules). To have the resulting DNA integrated into the host's chromosomal DNA as a provirus, where it remains (Reece et al., 2014). In 2014, 37 million people around the globe were estimated to live with the virus, 2 million new infections were reported and 1.2 million people died of AIDS related causes (UNAIDS, 2015b). The United Nations has estimated that more than 25 million people have lost their lives to the virus since its initial discovery in 1981, making it one of the highest mortality pandemics in recorded history. The disease gained a lot of attention in the early 80s, when large quantities of gay men in the United States of America started to show HIV-related symptoms (UNAIDS, 2015c). Nonetheless, the disease did not remain a local phenomenon for long and was within years classified as a world-wide pandemic. HIV is divided in two main types; HIV-1 and HIV-2 (Sharp and Hahn, 2011), Within the HIV-1 group, the main group (M), which is the pandemic form, has caused at least 25 million deaths (Merson et al., 2008). HIV-1 type M consists of nine distinct subtypes (A-D, F-H, J, K) and more than forty different circulating recombinant forms (CRFs) (Buonaguro et al., 2007; Goudsmit, 1997; Sharp and Hahn, 2011). The year 1981 is commonly considered as the birth-year of the modern-era HIV pandemic, but infections did occur earlier on. The oldest discovered HIV-1B RNA sequence stems from 1959 in the Democratic Republic of Congo (Zhu et al., 1998). Studies of the genetic structure and variation of suggest that the human variant of the virus is even older (Holmes, 2001). But pinpointing the exact 'birth-date' of the virus has proved a debatable matter. However, current consensus holds that initial infections were zoonotic; through multiple distinct transfers from chimpanzees carrying the simian variant of the virus; SIV, in sub-Saharan Africa during the early 20th century (Gao et al., 1999; Worobey et al., 2008). Based on the genetic variety that can be observed among the different viral types that reside in humans, it is likely that the virus was present in sub-Saharan Africa on a much smaller scale for a longer period of time, presumably with infected individuals being misdiagnosed with other diseases (Holmes, 2001).

Similarly to HCV, HIV is well suited for coalescent analysis (Pybus et al., 1999),

Methods

This section aims to explain the models of the two different Bayesian Coalescent Inference techniques included in this study.

Bayesian Skyline Plot

The relatively less intricate model; the BSP, which was introduced in Drummond et al. (2005) paper is explained here in detail. The BSP model is based on the Generalized Skyline Plot model (Strimmer and Pybus, 2001), which in turn is based on the initial Skyline Plot (Oliver G. Pybus, 2000).

Given input data of n contemporaneous sequences we can define $n-1$ times at which coalescent events could occur: $\boldsymbol{u} = \{u_1, u_2, ..., u_{n-1}\}$, i.e. $n-1$ times the phylogeny can change. During each coalescent interval $\Delta u_i = u_i - u_{i-1}$ we represent the number of lineages with a variable k_i The original Skyline Plot model makes direct use of these coalescent event in its calculations, with the assumption that the genealogy always changes at a coalescent event. This restriction resulted in noisy graphs and was revisited in the Generalized Skyline plot, which instead looks at possible groups of coalescent events. This leads to the introduction of the subset of group sizes $A = \{a_1, a_2, ...a_m\}$ subject to $(a_i > 0 \textit{ and } \sum_{i=1}^{m} a_i = n-1)$ that represents the number of coalescent events in each grouped interval, where $1 \leq m \leq n-1$ is the number of grouped intervals. The time at the end of each grouped interval is represented by the vector $w = \{w_1, w_2, ...w_m\}$ and the time spanned by each grouped interval is given by $\Delta w_j = w_j - w_{j-1}$.

The variables A, Θ and g, where $\Theta = \{\theta_1, \theta_2, ..., \theta_m\}$ is the effective population size for each grouped interval and g is the genealogy, define a piecewise demographic history with $2m-1$ demographic parameters and $n-1$ coalescent time parameters. In previous methods the genealogy was estimated/reconstructed a priori, a design choice that resulted in the model having to deal with the error associated with phylogenetic reconstruction. This opens up the possibility of heavily biased results for data sets with a sufficiently low variability (Drummond et al., 2005). The key idea of the BSP is that the genealogy should be sampled rather than estimated to remove said error. The log likelihood $f_G(g \mid \Theta, A)$

of the sampled genealogy given the demographic parameters is given by equation 1.

$$log f_G(g \mid \Theta, A) = \sum_{i=1}^{n+s-2} I_c(I) log \frac{k_i(k_i-1)}{2\theta_{h(i)}} - \frac{k_i(k_i-1)\Delta u_i}{2\theta_{h(i)}} \quad (1)$$

The upper bound of $n + s - 2$ is given by the fact that if we have heterochronous genealogies where we have n sequences sampled at s different times, then there are $n+s-2$ sampled intervals in total. The function $I_c(i)$ is an indicator function that tells us whether the ith event is a coalescent event. If it is, $I_c(i) = 1$, otherwise $I_c(i) = 0$. The function $h(i)$ maps the elements of u to w as explained in equation 2, and the parameter m is chosen a priori.

$$h(i) := \begin{cases} 1 & \text{if } \sum_{j=1}^{i} I_c(j) \leq a_1 \leqslant 1 \\ j, & \text{if } \sum_{k=1}^{j-1} a_k < \sum_{j=1}^{i} I_c(j) \leq \sum_{k=1}^{j} a_k \end{cases} \quad (2)$$

To construct the model two more assumptions are necessary. Firstly, we assume that each effective population size θ_t is drawn from an exponential distribution with the mean set to the previous population size. Secondly, we add a scale-invariant prior on the first element to state that we assume that our priori belief is unrelated to changes in timescale (Jeffreys, 1946). The multivariate prior distribution on Θ is shown in equation 3.

$$f_\Theta \propto \frac{1}{\theta_1} \prod_{j=2}^{m} \frac{1}{\theta_{j-1}} exp(-\theta_j/\theta_{j-1}) \quad (3)$$

After this initial estimation, the work on the Markov Chain Monte Carlo (MCMC) method done in (Drummond et al., 2002) is extended to iteratively improve the model. As aforementioned, the model samples the genealogy g, together with the effective population size through time Θ, the subset of grouped sizes A as well as the nucleotide-substitution model Ω given the data D. For a contemporaneous data set D the posterior distribution is shown in equation 4.

$$f_{iso}(\Theta, A, \Omega, g \mid D, \mu) = \frac{1}{Z} Pr\{D \mid g, \mu\} f_G(g \mid \Theta, A) \times f_\Theta(\Theta) f_A(A) f_\Omega(\Omega) \quad (4)$$

For such data the scaling variable μ, which converts units scaling per site to units of time, must be estimated before the MCMC method is applied. If however, the data contains heterochronous samples μ can instead be sampled by the MCMC. This gives us equation 5.

$$f_{het}(\Theta, A, \Omega, g, \mu \mid D) = \frac{1}{Z} Pr\{D \mid g, \mu\} f_G(g \mid \Theta, A) \times f_\Theta(\Theta) f_A(A) f_\Omega(\Omega) \quad (5)$$

The use of the MCMC results in j states of the simulated parameters, and most importantly, j states of Θ. With this data it is possible to reconstruct the demographic history $\theta(t)$ though time. From this distribution the mean, median and the 95% highest posterior density intervals can be calculated for $\theta(t)$ at each time of interest t_i. The Bayesian Skyline Plot is constructed with these values as can be seen in the Results section.

Extended Bayesian Skyline Plot

This article is based on and aims to recreate the results in Drummond et al. (2005) and therefore the BSP model has been explained in detail. However, the state of the art EBSP is also studied. Although the model will not be explained to the same extent as the BSP, some description of the model will be given. The interested reader is welcome to venture to the Heled and Drummond (2008) paper for a more in depth explanation.

The EBSP is indeed quite similar in logic to the BSP, but extends the older method in several ways. It permits analysis of multiple loci, it uses Bayesian stochastic variable selection and supports piecewise linear demographic functions. While it is often relevant to study multiple loci, the method also enhances the performance of single-loci analysis since there is no longer a need to specify the number of groups (m) a priori. The single loci analysis has been carried out in this article. The log likelihood of the genealogy is given by equation 6.

$$log f_G\{g_k|\Theta, \Lambda\} = \sum_{i=n_k}^{2} log \frac{\binom{i}{2}}{\theta_k(u_k, i)} + \int_{u_k,i+1}^{u_k,i} \frac{\binom{i}{2}}{\theta_k(t)} \mathrm{d}t \quad (6)$$

Where the new parameter Λ is an indicator vector of values $\lambda \in \{0, 1\}$ that indicates if the corresponding value of Θ shall contribute to the demographic model. We again make the assumption that the effective population size follows an exponential distribution, here with a mean ϕ, and the priori distribution of Θ is given by equation 7.

$$f_\Theta(\Theta, \phi) = f_\phi(\phi) \prod_{j=1}^{n} \frac{1}{\phi} e^{-\theta_j/\phi} \quad (7)$$

The posterior distribution of the MCMC sampled by the EBSP is given by equation 8.

$$f(\Theta, \Lambda, g, \phi, \mu \mid D, P) = \frac{1}{Z}\left[\sum_{k=1}^{m} f_D\{D_k \mid g_k, \mu\} f_G(g_k \mid \Theta, \Lambda)\right] f_\Theta(\Theta, \phi) f_\Lambda(\Lambda) f_\mu(\mu) \quad (8)$$

Where $P = \{p_1, p_2, ..., p_k\}$ is a population size factor of g_k. This parameter accounts for any differences in ploidy and/or mode of inheritance among loci.

Data

Specific data sequences were employed in order to carry out analyses, these can have a big impact on the results' accuracy. The HCV dataset comprises of 63 partial gene sequences of type 4 and 5 partial sequences of type 1, subtype G and is the exact same data set as used in Drummond et al. (2005). The HIV-1 subtype B dataset was composed by manual selection from the Los Alamos HIV online database. Full-genome sequences were used in all the HIV analyses.

To confirm the validity of the used samples (i.e. all samples residing in the B subtype) used in 4, a phylogenetic tree was constructed, in combination with a newly introduced outgroup (i.e. a less closely related organism) an HIV-1 subtype C sequence, using RAxML V.8.2.4 (Stamatakis, 2014). The premise of this approach is that the subtype B sequences should be more closely related and subtype C should thus be an outgroup in the resulting tree (this approach is fairly conventional when determining phylogeny, see e.g. Lemey (2009)). The visualization to enable visual inspection and rooting of the tree were achieved using FigTree (Rambaut, 2007). To align the sequences ClustalW (Thompson et al., 2002) was used. Subtype C did indeed not cluster with the sequences from subtype C, as can be observed in figure 6. We can thus infer that all used employed sequences are more closely related to each other than a presumed phylogenetic close organism, which is consistent with all the sequences belonging to subtype B.

Results

Part of the analysis involved the use of available open software. The study of the Bayesian Skyline Plot was aided by the use of the BEAST 1.4 software and the corresponding version of BEAUTi. For the extended version of the technique (EBSP) the latest versions were put to use, i.e. BEAST and BEAUTi version 1.8.3. Figures were generated by Tracer v. 1.6 and the Python matplotlib library (Hunter et al., 2007). Some instances required a MCMC chain length of exceptional length (up to 200 million) to generate an effective sample size (ESS) of adequate size[2] (> 100), which requires considerable computational power. Thus the CIPRES V.3.3 cluster (Miller et al., 2010) was used to compute a portion of the experiments.

In Drummond et al. (2005) the authors used an exponential distribution with a mean proportional to the previous effective population size. However, following this approach appeared to be infeasible due to software constraints. Thus, henceforth the mean for all examples (unless stated otherwise) is equal to 1.0. The variables for each experiment were sampled every 1000 iterations. Drummond et al. Drummond et al. reached convergence of their MCMC for the HCV samples after merely 10 million iterations . In our study however, 10 million iterations only resulted in an ESS of 10 for the EBSP on the same data set. The experiment was repeated for a chain length of 200 million, which resulted in a more satisfactory ESS of 137. The resulting skyline plot can be seen in figure 1. Since the BSP samples fewer variables it only required 50 million iterations for the same experiment, of which the result can be seen in figure 2

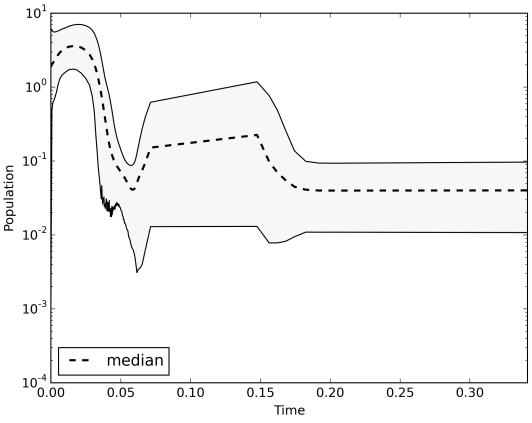

Figure 1: Relative population size of the HCV virus in Egypt through time generated with the Extended Bayesian Skyline Plot after a Markov chain length of 200 000 000 with an exponential distribution. Effective sample size is 137. The Y-axis shows the relative population size and the X-axis shows the number of years before 1993 (shown in thousands of years). The thick dotted line shows the median value, and the thinner whole-drawn lines shows the 95% highest posterior density limits.

Other priori distributions and parameters were also tested but resulted in less valid results. For example, one simulation was ran when the a priori nucleotide-substitution model set to GTR instead of the original HKY and some priori distributions were changed from uniform to normal. This run did not generate an ESS higher than 90, even after 250 million iterations and was therefore excluded from the rest of the study (skyline plot not shown).

For comparison, the skyline plot showing the effective population size of the HCV virus in Egypt from Heled and

[2]If the Markov Chain's effective sample size is too low, the chain contains a lot of correlated samples, which might bias the posterior distribution.

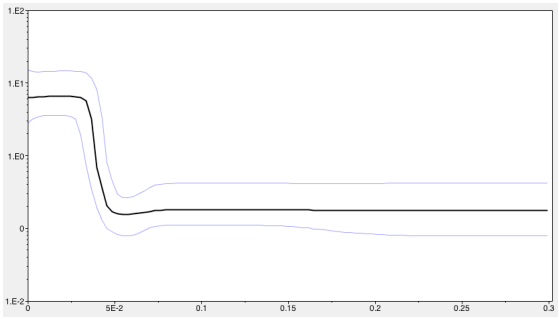

Figure 2: Relative population size of the HCV virus in Egypt through time generated with the Bayesian Skyline Plot after 50 000 000 iterations with an exponential distribution. Effective sample size is 134. The Y-axis shows the relative population size and the X-axis shows the number of years before 1993 (shown in thousands of years). The thick dotted line shows the median value, and the thinner whole-drawn lines shows the 95% highest posterior density limits.

Drummond (2008) has been included in this paper, which gives a comparison of the BSP and the EBSP. The result can be seen in figure 3.

The EBSP has also been applied to samples of HIV-1 subtype B. Figure 4 shows the relative population size through time for said virus, based on samples from North America, Europe and sub-Saharan Africa. The resulting ESS was a satisfactory 3844. The simulation was carried out anew, with a subset of the samples, only including sequences from the United States, which resulted in figure 5.

Discussion

As aforementioned in the methods section the Extended Bayesian Skyline model has several advantages over the previous version, with one of the most prominent in single-loci analysis being that the hyper parameter m does not have to be set a priori. The authors of Heled and Drummond (2008) do in fact argue that the value of $m = 24$ as set here and in Drummond et al. (2005) for the HCV plots is in fact too high which might lead to biased results. The authors also argue that the performance of the EBSP is better than that of the BSP, i.e. that better results with an higher ESS are obtained faster. We did however find results pointing to the opposite. We had to use a chain length of 4 times the size of that in the BSP when running the same samples through the EBSP model to generate an ESS equal to that of the BSP. The computation time needed is directly influenced by the length of the Markov chain. However, due to time constraints, not all settings and parameters were explored exhaustively for either model.

The initial population size has not been set to real values for any of the generated plots. This has resulted in relative results (i.e. scaling factors) instead of the exact values found

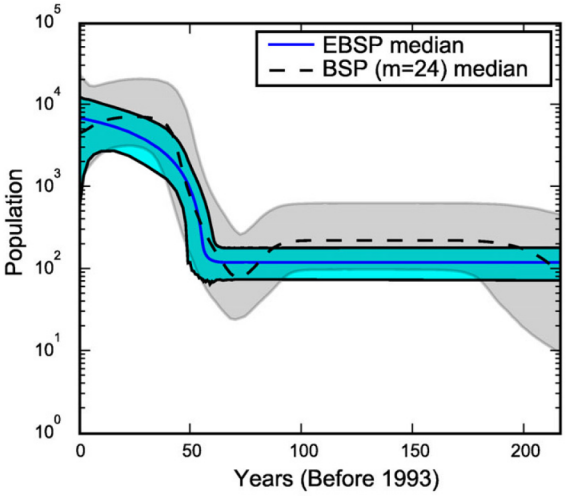

Figure 3: Effective population size of the HCV virus in Egypt through time, plotted with the Extended Bayesian Skyline Plot from (Heled and Drummond, 2008) compared to results generated with the BSP for the same samples. The blue plot shows the demographic history generated by the EBSP while the gray plot shows the corresponding plot for the BSP with $m = 24$. The thick solid/dashed lines in the middle of the two plots show the median value, and the thinner dashed lines show the 95% highest posterior density limits. The Y-axis shows the effective population size and there X-axis shows number of years since 1993.

in the original paper. The figure associated with the population size in the USA, figure 5, show some surprising values for the X-axis. And while we are unsure what caused this, we would speculate that for heterochronous samples they are too close (chronologically) to each other (samples from 2012 to 2014). The conversion parameter μ could thus have been inaccurately sampled, leading to an inaccurate timescale. The trend still substantiates our expectations, so given more chronologically diverse samples we expect a correct scale. This hypothesis is reinforced by the satisfactory results when a more heterochronous sample set (dating from 2014 to the early 1980s) was used in figure 4.

HCV

The results obtained from EBSP on the HCV samples can be seen in figure 1 and the corresponding results from the BSP in figure 2. They seem to mirror the results achieved in the original papers quite well, see figure 3. The graphs are not identical, probably due to the fact that not all parameters were specified in the original paper and thus some exploring was necessary on our part. The general trend does seem to be the same for our results and those of the original authors which leads us to believe that even though our results are not identical to those in the original papers, they are acceptable in the context of this article.

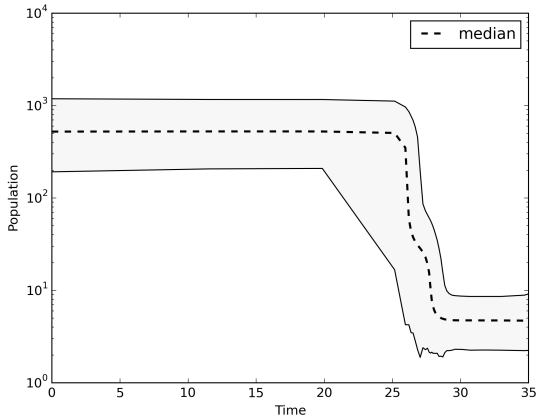

Figure 4: Relative population size through time of the HIV-1 subtype B virus, generated by the EBSP with samples from North America, Europe and sub-Saharan Africa. The X-axis shows years before 2014. Estimated sample size is 3844. The Y-axis shows the relative population size and the X-axis shows the number of years before 2014. The thick dotted line shows the median value, and the thinner whole-drawn lines shows the 95% highest posterior density limits.

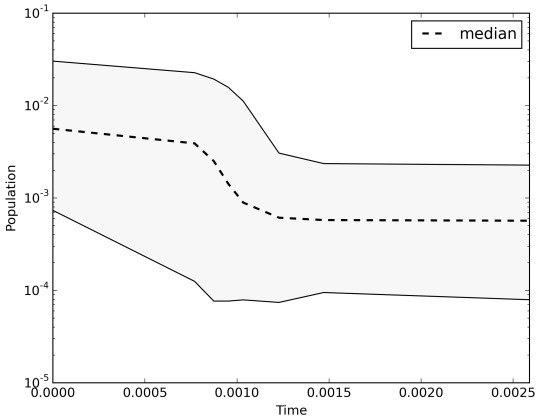

Figure 5: Relative population size through time of the HIV-1 subtype B virus, generated by the EBSP with samples taken from the United States of America. The Y-axis shows the relative population size and the X-axis shows the amount of time that has passed since 2014 (scale unknown or inaccurately generated). The thick dotted line shows the median value, and the thinner whole-drawn lines shows the 95% highest posterior density limits.

HIV-1

Our results regarding the HIV experiments seem to corroborate the real world scenario adequately. Both the dramatic increase in infections in the 80s and the low population size before that period are visible, see figure 4. Since its initial surge in the 80s the prevalence of the virus in the world steadily increased until its peak in the beginning of the 21th century, and has proportionally declined hitherto, especially in the west (UNAIDS, 2015a). It thus seems feasible to consider the mean of the effective population size as relatively constant, as can be seen in figure 4.

The demographic behaviour of HIV-1 subtype B in the United States is especially interesting, due to its initially very local propagation throughout clusters of gay men in California and New York (Kuiken et al., 2000). Maximum likelihood and Bayesian Inference techniques suggest that the spread to the Americas occurred through a single introductory event in the Caribbean in the early 1960s (Junqueira et al., 2011). The initial continent crossing most likely occurred with an Haitian worker who return from work missions in Congo. Onward, the virus rapidly disseminated to the rest of the the Americas, including the United States. Figure 5 shows this rapid growth in the effective population size of the virus in the USA. AIDS is a sexually transmittable disease and it has been shown that a scale-free network can accurately capture the contamination scheme of a community (Schneeberger et al., 2004). Should the virus find and infect a so-called hub (i.e. a well connected individual) in such a network, it is likely that the infectious disease will disseminate quickly throughout the entire network. This offers an explanation for the remarkable rapid growth of the virus in the United States, where an early infected patient, commonly referred to as patient 0, was reported to have had approximated 750 homosexual men as sexual partners between 1979 and 1981 (Auerbach et al., 1984). These findings could account for HIV-1B's rapid spread in the homosexual community in the country. Our results fit this scenario very well; we can observe a swift increase in the effective population size of the virus around that time, see figures 4 and 5.

While HIV-1B is by far the most prevalent type of the virus found in the western world (Kuiken et al., 2000) there are a myriad of other subtypes and recombinant forms. These have however not been included in this study, but they could be investigated analogously.

Acknowledgments

We would like to thank professor Oliver Pybus, who made the original dataset from Drummond et al. (2005) available to us. We would also like to thank Pieter Libin who gave us the inspiration which led us to pursue the line of work presented in this article. Furthermore we would like to express our gratitude to the people behind the CIPRES Science Gateway (Miller et al., 2010), the Hydra cluster. and the Los Alamos National Laboratory for making computing power and sequence datasets available.

References

Auerbach, D. M., Darrow, W. W., Jaffe, H. W., and Curran, J. W. (1984). Cluster of cases of the acquired immune deficiency syndrome: patients linked by sexual contact. *The American journal of medicine*, 76(3):487–492.

Buonaguro, L., Tornesello, M., and Buonaguro, F. (2007). Human immunodeficiency virus type 1 subtype distribution in the worldwide epidemic: pathogenetic and therapeutic implications. *Journal of virology*, 81(19):10209–10219.

Desport, M. (2010). *Lentiviruses and macrophages: molecular and cellular interactions*. Horizon Scientific Press.

Donnelly, P. and Tavare, S. (1995). Coalescents and genealogical structure under neutrality. *Annual review of genetics*, 29(1):401–421.

Drummond, A. J., Nicholls, G. K., Rodrigo, A. G., and Solomon, W. (2002). Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. *Genetics*, 161(3):1307–1320.

Drummond, A. J., Rambaut, A., Shapiro, B., and Pybus, O. G. (2005). Bayesian coalescent inference of past population dynamics from molecular sequences. *Molecular biology and evolution*, 22(5):1185–1192.

Gao, F., Bailes, E., Robertson, D. L., Chen, Y., Rodenburg, C. M., Michael, S. F., Cummins, L. B., Arthur, L. O., Peeters, M., Shaw, G. M., et al. (1999). Origin of hiv-1 in the chimpanzee pan troglodytes troglodytes. *Nature*, 397(6718):436–441.

Goudsmit, J. (1997). *Viral sex: the nature of AIDS*. Oxford University Press on Demand.

Griffiths, R. C. and Tavare, S. (1994). Sampling theory for neutral alleles in a varying environment. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 344(1310):403–410.

Harding, R. M. (1996). New phylogenies: an introductory look at the coalescent. *New Uses for New Phylogenies*, pages 15–22.

Heled, J. and Drummond, A. J. (2008). Bayesian inference of population size history from multiple loci. *BMC Evolutionary Biology*, 8(1):289.

Holmes, E. C. (2001). On the origin and evolution of the human immunodeficiency virus (hiv). *Biological Reviews of the Cambridge Philosophical Society*, 76(02):239–254.

Hunter, J. D. et al. (2007). Matplotlib: A 2d graphics environment. *Computing in science and engineering*, 9(3):90–95.

Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. In *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, volume 186, pages 453–461. The Royal Society.

Joy, D. A., Feng, X., Mu, J., Furuya, T., Chotivanich, K., Krettli, A. U., Ho, M., Wang, A., White, N. J., Suh, E., et al. (2003). Early origin and recent expansion of plasmodium falciparum. *science*, 300(5617):318–321.

Junqueira, D. M., De Medeiros, R. M., Matte, M. C. C., Araújo, L. A. L., Chies, J. A. B., Ashton-Prolla, P., and de Matos Almeida, S. E. (2011). Reviewing the history of hiv-1: spread of subtype b in the americas. *PloS one*, 6(11):e27489.

Kingman, J. F. (1982a). On the genealogy of large populations. *Journal of Applied Probability*, pages 27–43.

Kingman, J. F. C. (1982b). The coalescent. *Stochastic processes and their applications*, 13(3):235–248.

Kuiken, C., Thakallapalli, R., Eskild, A., and de Ronde, A. (2000). Genetic analysis reveals epidemiologic patterns in the spread of human immunodeficiency virus. *American journal of epidemiology*, 152(9):814–822.

Lavanchy, D. (2011). Evolving epidemiology of hepatitis c virus. *Clinical Microbiology and Infection*, 17(2):107–115.

Lemey, P. (2009). *The phylogenetic handbook: a practical approach to phylogenetic analysis and hypothesis testing*. Cambridge University Press.

Merson, M. H., O'Malley, J., Serwadda, D., and Apisuk, C. (2008). The history and challenge of hiv prevention. *The Lancet*, 372(9637):475–488.

Miller, M. A., Pfeiffer, W., and Schwartz, T. (2010). Creating the cipres science gateway for inference of large phylogenetic trees. In *Gateway Computing Environments Workshop (GCE), 2010*, pages 1–8. IEEE.

Oliver G. Pybus, Andrew Rambaut, P. H. H. (2000). An integrated framework for the inference of viral population history from reconstructed genealogies. *GENETICS*, 155(3):1429–1437.

Pybus, O. G., Holmes, E. C., and Harvey, P. H. (1999). The mid-depth method and hiv-1: a practical approach for testing hypotheses of viral epidemic history. *Molecular biology and evolution*, 16(7):953–959.

Rambaut, A. (2007). Figtree v1.4.2.

Ray, S. C., Arthur, R. R., Carella, A., Bukh, J., and Thomas, D. L. (2000). Genetic epidemiology of hepatitis c virus throughout egypt. *Journal of Infectious Diseases*, 182(3):698–707.

Reece, J. B., Campbell, N. A., Urry, L. A., Cain, M. L., Wasserman, S. A., Minorsky, P. V., and Jackson, R. B. (2014). *Introduction to Viruses.*, book section 26, pages 621–627. Pearson Education Limited.

Schneeberger, A., Mercer, C. H., Gregson, S. A., Ferguson, N. M., Nyamukapa, C. A., Anderson, R. M., Johnson, A. M., and Garnett, G. P. (2004). Scale-free networks and sexually transmitted diseases: a description of observed patterns of sexual contacts in britain and zimbabwe. *Sexually transmitted diseases*, 31(6):380–387.

Sharp, P. M. and Hahn, B. H. (2011). Origins of hiv and the aids pandemic. *Cold Spring Harbor perspectives in medicine*, 1(1):a006841.

Stamatakis, A. (2014). Raxml version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, page btu033.

Strickland, G. T. (2006). Liver disease in egypt: hepatitis c superseded schistosomiasis as a result of iatrogenic and biological factors. *Hepatology*, 43(5):915–922.

Strimmer, K. and Pybus, O. G. (2001). Exploring the demographic history of dna sequences using the generalized skyline plot. *Molecular Biology and Evolution*, 18(12):2298–2305.

Thompson, J. D., Gibson, T., Higgins, D. G., et al. (2002). Multiple sequence alignment using clustalw and clustalx. *Current protocols in bioinformatics*, pages 2–3.

UNAIDS (2015a). Aids by the numbers.

UNAIDS (2015b). Fact sheet.

UNAIDS (2015c). How aids changed everything.

Worobey, M., Gemmel, M., Teuwen, D. E., Haselkorn, T., Kunstman, K., Bunce, M., Muyembe, J.-J., Kabongo, J.-M. M., Kalengayi, R. M., Van Marck, E., et al. (2008). Direct evidence of extensive diversity of hiv-1 in kinshasa by 1960. *Nature*, 455(7213):661–664.

Zhu, T., Korber, B. T., Nahmias, A. J., Hooper, E., Sharp, P. M., and Ho, D. D. (1998). An african hiv-1 sequence from 1959 and implications for the origin of the epidemic. *Nature*, 391(6667):594–597.

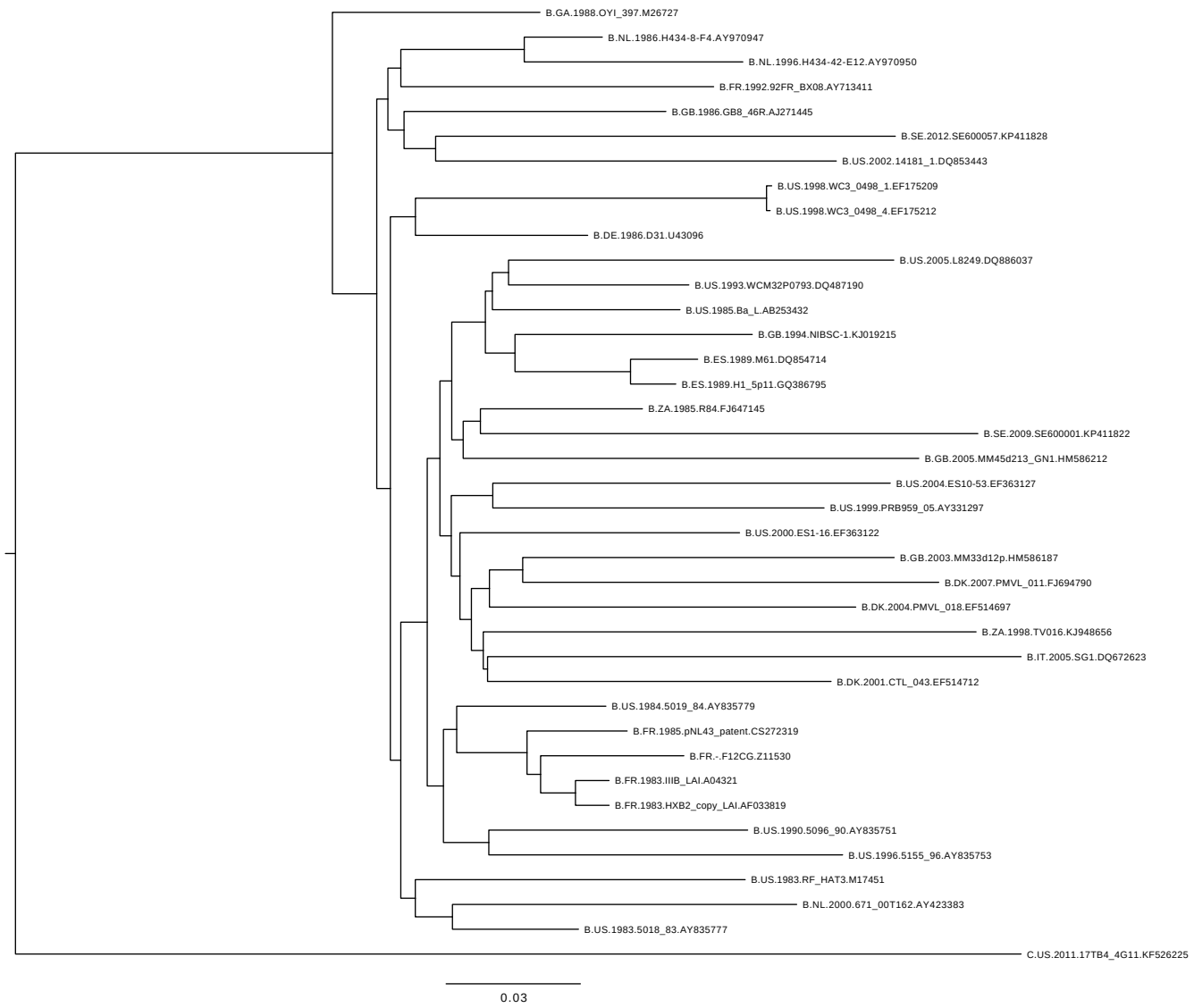

Figure 6: Phylogenetic tree of HIV-1 subtype B samples used in this study as well as an HIV-1C sample. Clearly not clustering within the subtype B samples and thus serving as an outgroup.