PIQMEE: Bayesian Phylodynamic Method for Analysis of Large Data Sets with Duplicate Sequences

Veronika Boskova[1,2,3,*] and Tanja Stadler[1,2]

[1]Department of Biosystems Science and Engineering, ETH Zürich, Basel, Switzerland

[2]Swiss Institute of Bioinformatics (SIB), Switzerland

[3]Center for Integrative Bioinformatics Vienna, Max Perutz Labs, University of Vienna and Medical University of Vienna, Vienna, Austria

*Corresponding author: E-mail: boskovav57@univie.ac.at.

Associate editor: Thomas Leitner

Abstract

Next-generation sequencing of pathogen quasispecies within a host yields data sets of tens to hundreds of unique sequences. However, the full data set often contains thousands of sequences, because many of those unique sequences have multiple identical copies. Data sets of this size represent a computational challenge for currently available Bayesian phylogenetic and phylodynamic methods. Through simulations, we explore how large data sets with duplicate sequences affect the speed and accuracy of phylogenetic and phylodynamic analysis within BEAST 2. We show that using unique sequences only leads to biases, and using a random subset of sequences yields imprecise parameter estimates. To overcome these shortcomings, we introduce PIQMEE, a BEAST 2 add-on that produces reliable parameter estimates from full data sets with increased computational efficiency as compared with the currently available methods within BEAST 2. The principle behind PIQMEE is to resolve the tree structure of the unique sequences only, while simultaneously estimating the branching times of the duplicate sequences. Distinguishing between unique and duplicate sequences allows our method to perform well even for very large data sets. Although the classic method converges poorly for data sets of 6,000 sequences when allowed to run for 7 days, our method converges in slightly more than 1 day. In fact, PIQMEE can handle data sets of around 21,000 sequences with 20 unique sequences in 14 days. Finally, we apply the method to a real, within-host HIV sequencing data set with several thousand sequences per patient.

Key words: **Bayesian phylodynamics, duplicate sequences, subsampling, large data sets, BEAST 2, fast algorithms.**

Introduction

Phylogenetic and phylodynamic studies of pathogen spread at both between- and within-host scales rely on genetic sequence data as input. For estimating the between-host dynamics, a single (usually consensus) pathogen sequence per patient is often sufficient for conducting successful analyses (Drummond et al. 2005; Stadler et al. 2013, 2014; Volz and Pond 2014; Faria et al. 2016). The usual data set size in such studies is in the order of tens to hundreds of sequences.

Many pathogens, such as RNA viruses, replicate and mutate very quickly within a host (Hué et al. 2005; Pybus and Rambaut 2009; Alizon and Fraser 2013). By consequence, within a very short time such pathogens create an entire population of a virus called quasispecies (Eigen and Schuster 1977; Wilke 2005; Domingo et al. 2012), characterized by high sequence duplicity and diversity within a single host (Boeras et al. 2011; Domingo et al. 2012; Töpfer et al. 2014; Wu et al. 2014). In the context of viral quasispecies, each unique sequence is referred to as a haplotype (Töpfer et al. 2013).

The sequences from the quasispecies population are usually obtained using either labor-intensive cloning combined with Sanger sequencing or faster and more efficient next-generation sequencing (Schuster 2008; Goodwin et al. 2016). Next-generation sequencing, however, produces short reads that require further processing, whereby the reads are stitched together in order to reconstruct the original sequence (Goodwin et al. 2016). Assuming we can overcome, or at least correct for, the known errors of the RNA amplification procedure (McKinley et al. 2011), sample preprocessing (Vrancken et al. 2016), and data postprocessing errors (Beerenwinkel et al. 2012), we can reconstruct the within-host pathogen diversity from next-generation sequencing data sets in great detail (Zagordi et al. 2011; Schirmer et al. 2014; Prosperi et al. 2013; Töpfer et al. 2013, 2014; Pandit and de Boer 2014; Malhotra et al. 2016). Depending on the depth of the sequencing and performance of the assembly method, next-generation sequencing followed by haplotype reconstruction yields data sets of tens to hundreds of haplotypes each in multiple copies, annotated as a haplotype frequency.

Article

The reconstructed haplotypes are then aligned to create a multiple sequence alignment. Before it can be used as input for current phylogenetic and phylodynamic inference methods, this alignment needs to be re-expanded, each unique haplotype being duplicated many times, proportional to its frequency. The phylogenetic inference methods then produce a phylogenetic tree, on which all sequences in the alignment are represented as tips. However, when re-expanded, the sequence alignment contains thousands of sequences. Reconstructing the phylogeny and the population dynamics from such big data sets represents a substantial computational hurdle.

Maximum-likelihood (ML) approaches have been especially useful to treat such large data sets (Stamatakis 2014; Montoya et al. 2016; Minh et al. 2020). However, the disadvantage of these methods is that they first maximize the likelihood over different phylogenetic trees, heuristically searching tree space to find the best-fitting tree. Then, given this ML tree, the pathogen dynamics are inferred. During tree inference, identical sequences will be grouped together, and because the most likely divergence between two identical sequences is always 0, the branch lengths of these subtrees will be (very close to) 0 or set to some default minimum (e.g., 10^{-6} when using RAxML [Stamatakis 2014]). This means that in the ML framework, the tree will represent an extreme phylogeny that does not correspond to the real history of the sampled sequences. As the ML procedure is based on inferring a single ML tree first, the bias introduced by this extreme topology will propagate to all subsequent phylodynamic analyses.

In contrast, Bayesian phylodynamic methods provide a natural way of integrating over phylogenetic trees, τ, when inferring population dynamic parameters, η, and sequence evolution parameters, θ. These methods infer the joint posterior distribution of τ, η, and θ given the data, $f[\tau, \theta, \eta|\mathrm{D}]$, where D is a sequence alignment. When a parameter is not one wants to integrate over the uncertainty of a given parameter, it is possible to marginalize the posterior distribution over this so-called nuisance parameter.

The posterior distribution is calculated as the product of the phylogenetic likelihood, $f[\mathrm{D}|\tau, \theta]$, the tree prior (also called the phylodynamic likelihood), $f[\tau|\eta]$, and the parameter prior distributions, $f[\eta]$ and $f[\theta]$, and is normalized by the marginal distribution of the data, $f[\mathrm{D}]$:

$$f[\tau, \theta, \eta|\mathrm{D}] = \frac{f[\mathrm{D}|\tau, \theta]f[\tau|\eta]f[\eta]f[\theta]}{f[\mathrm{D}]}. \quad (1)$$

As the denominator at the right-hand side of the equation is difficult to evaluate, classic phylodynamic methods rely on numerically approximating the posterior distribution by sampling from it. Sampling from the posterior distribution is performed using the Metropolis–Hastings Markov chain Monte Carlo (MCMC) procedure (Metropolis et al. 1953; Hastings 1970). However, this procedure is computationally expensive and is not well suited for data sets larger than a few hundred sequences (Poon et al. 2012).

Options to speed up the calculation of the posterior density have been proposed and implemented. One approach is to parallelize the phylogenetic likelihood $f[\mathrm{D}|\tau, \theta]$ calculation using multicore architecture of graphic processing units, as in the BEAGLE software (Suchard and Rambaut 2009; Ayres et al. 2019). Another alternative is using sequential Monte Carlo methods for sampling phylogenetic trees (Bouchard-Côté et al. 2012). Despite these advances, the MCMC methods are slow and often incapable of processing the full data set when the input sequence count is larger than a few hundred.

To address the issue of computational expense when applying Bayesian methods to large data sets with duplicate sequences, one of the following two approaches is usually employed. The first approach is to randomly subsample the full data set to keep the diversity but to decrease the computational burden (Poon et al. 2011, 2012). This approach should yield unbiased estimates of population dynamic model parameters, but may lead to inference of a most recent common ancestor (MRCA) that is younger than, and thus not representative of, the MRCA of the full data set. This happens if the random sample chosen does not contain the two most divergent sequences that define the MRCA of the full data set. Furthermore, as a lot of information is left out, precision of the parameter estimates is compromised if a subsample instead of the whole data set is used. The second way of reducing the computational burden in the Bayesian methods is to perform inference on only the unique sequences, completely ignoring their respective frequencies (Bull et al. 2011; Recarey and Cristina 2014). However, this amounts to inference on a biased data set and thus parameter estimates may be biased as well. To our knowledge, these biases have not been explored previously.

In this study, we show that analyses with only unique sequences or only randomly subsampled data sets result in less accurate and/or less precise parameter estimates than when all data are being used. More importantly, we present a new method that improves convergence of the MCMC by keeping track of the duplicate sequence branching times but not of the full topology formed by these duplicate sequences. The rationale for this is that the duplicate sequences provide information regarding the evolutionary rate and thus the branching times, and can further help refine the topology on the unique sequences (DeWitt et al. 2018). However, these same sequences provide no information to narrow down the plausible tree topology space of the duplicate sequence subtrees (Dudas and Bedford 2019), making the tracking of the duplicate sequence subtree topology unnecessary. We implement this method as a BEAST 2 (Bouckaert et al. 2014) package called PIQMEE, which stands for "phylogenetic inference of quasispecies molecular evolution and epidemiology." The method works with any sequence evolution model currently available in BEAST 2. As the population dynamics model, we adapted the birth–death skyline (BDsky) model (Stadler et al. 2013). We show on simulations that the PIQMEE method is accurate and precise, and faster than the classic method when the alignment analyzed contains duplicate sequences. Finally,

we apply our method to an empirical HIV data set containing thousands of sequences.

New Approaches

The main idea of the new approach is based on the observation that a set of identical sequences contains no information regarding the underlying tree topology. Sampling various topological configurations of the subtree of identical sequences in MCMC is therefore a waste of computational time. Thus, especially for large data sets with many duplicate sequences, when including all sequences and representing each sequence as a tip in the tree, the MCMC chains mix very slowly. This then needs to be compensated for by simulating long chains. A lot of computational effort could be saved if, during MCMC, the topology of identical sequences was not sampled, meaning that the duplicate sequences were not treated as separate tips in the phylogenetic tree. However, the duplicates cannot be completely ignored, because such an approach would lead to biases.

We therefore propose calculating the posterior probability of the tree and the parameters for a reduced tree structure (see the Materials and Methods section for details). The reduced phylogenetic tree is built only from unique sequences but is complemented by an array of branching and sampling times of all the duplicate sequences.

The main assumption of the method is that the duplicate sequences of the haplotype always arise from an already existing sequence through branching (duplication) and not through mutation. In other words, we allow each haplotype to arise only once during the population history. If we were to represent this assumption on a full, nonreduced tree, it would lead to a tree which is "recursively monophyletic" with respect to the identical sequences. In such a recursively monophyletic tree, there is always at least one monophyletic group of identical sequences. If we were to remove it from the tree, there would be another monophyletic group of identical sequences formed. One could continue with the cycle of identification and removal of monophyletic groups of identical sequences until no tips would be left in the tree.

We adapted the calculation of the phylogenetic likelihood $f[\mathrm{D}|\tau,\theta]$ and the tree prior $f[\tau|\eta]$ for this reduced tree representation (see the Materials and Methods section for details) and implemented the tree structure as well as the new likelihood functions in a BEAST 2 package PIQMEE. The source code and the executable jar file can be obtained from https://github.com/boskovav/piqmee (last accessed May 28, 2020).

Results

Impact of Input Number of Sequences on Method Performance

We simulated 100 trees for each of the following tip counts: 300, 1,200, 2,100, 3,000, and 6,000 tips. All tips were sampled at the same time point. The tree model used was a birth–death model with a constant birth and death rate. We simulated sequences on these trees according to a Jukes–Cantor model with a strict clock (for details, see the Materials and Methods section). We analyzed the resulting sequences using four inference methods. First, we analyzed the full data set with our new method, which we will refer to in the results as PIQMEE_all. Second, we analyzed the full data set with the classic BEAST 2 method, referred to as CLASSIC_all. Third, we analyzed only the unique sequences using the classic method (CLASSIC_unique). Fourth, we randomly subsampled the full data set to the size equal to the number of unique sequences and analyzed this subset with the classic method (CLASSIC_random). The BDsky model was assumed for the tree prior in all four analyses.

The results of the analyses are shown in figure 1 and table 1. Analyses of all data sets ran well with the exception of 68 data sets with 6,000 sequences in the CLASSIC_all method. These runs never started due to the BEAST 2 Java application having insufficient memory to handle the amount of data loaded. The remaining 32 runs ran for 7 days but only 20 reached an effective sample size (ESS) of 200 for all the metrics and model parameters. This shows that data sets of this size are at the limit of what can be handled by the software, if each duplicate sequence corresponds to a separate tip in the tree.

Table 1 and figure 1 clearly show that the analyses of subsets of the full data, that is, unique or random subsets of sequences, are much faster than the analyses of the full data set under either the CLASSIC or PIQMEE methods. Note that as the number of sequences in the unique and random data sets stays more-or-less constant (median of 19–20 sequences), these analyses always take approximately the same amount of time no matter how many sequences there are in the full data set.

In what follows, we use the root mean squared error (RMSE) of the median parameter estimates as a measure of accuracy. Systematic deviation from the true value calculated as "MoM—true value", where MoM stands for the median of medians estimates, is used as a measure of bias. Precision refers to the width of the 95% highest posterior density (HPD) interval.

First, we will discuss the results obtained from the CLASSIC_unique method. Analyses of the unique sequences lead to biased parameter estimates as indicated by the systematic deviation of the MoM from the true values of the parameters. The posterior estimates of the tree height parameter are above the true parameter value. This overestimation likely happens because in the unique sequence tree only the branches with mutation are seen. By consequence, at the considered substitution rate the plausible trees have longer branches than the true tree in order to accommodate all this "quickly appearing" diversity. The population dynamic parameter R_e (effective reproductive number), which is the birth rate divided by the death rate, is biased upwards and the death rate is biased downwards. The bias in the population dynamic parameters could be purely a consequence of the overestimation of the tree height. However, it could also be explained by the lack of short terminal branches in the tree, which constitute a significant proportion of all branches leading to tips in our simulations. These two hypotheses can be probed by fixing the tree and reestimating the parameters (see supplementary fig. S1, Supplementary Material online).

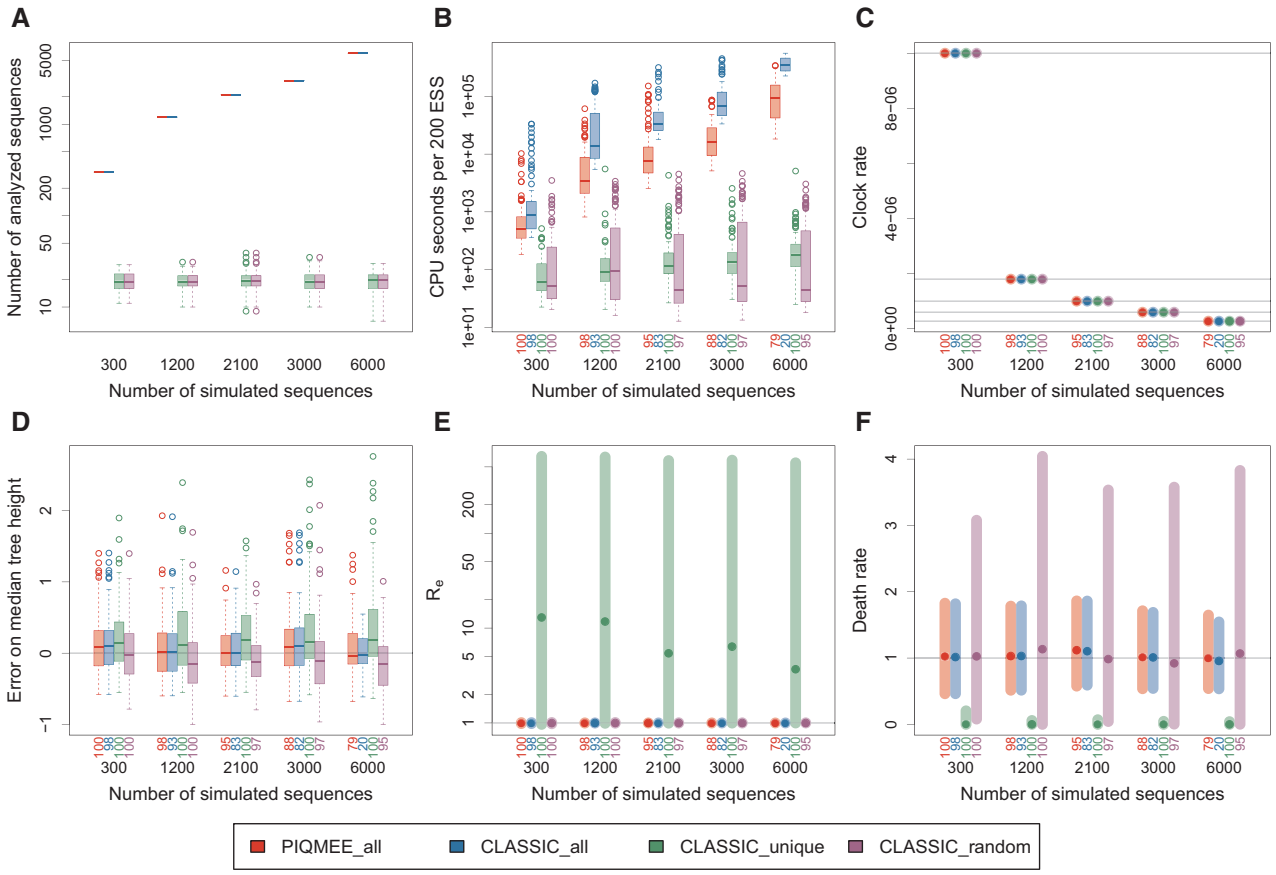

Fig. 1. Performance comparison of the PIQMEE method and the CLASSIC method on full, unique and randomly subsampled data sets of various sizes. For each setting, 100 data sets were analyzed. All the summary statistics shown are only for the runs that reached ESS of 200 for each parameter. The total number of such runs out of 100 is shown in smaller, colored numbers on the *x*-axis below each figure. (*A*) The number of sequences analyzed by each method. For PIQMEE_all and CLASSIC_all, the method was considering all the sequences. For CLASSIC_unique and CLASSIC_random, the size of the data set was smaller. (*B*) The distribution of the CPU seconds elapsed until the runs reached ESS of 200 for all the parameters. (*C*) The clock rate used for inference, in units of substitutions/site/time unit. As all our data sets have the sequences sampled at one time point, we need to fix the clock rate to infer the tree height. The clock rate for all analyses was fixed to the values used for simulations. (*D*) The distribution of the normalized relative error of median posterior estimates of the tree heights ($\frac{\text{median estimate}-\text{true}}{\text{true}}$). The ideal case of error being 0 is indicated by the gray line. (*E*) and (*F*) The distribution of the effective reproductive number ($R_e = \frac{\text{birth rate}}{\text{death rate}}$) and the death rate estimates. The top and the bottom of the bars represent the median value of the top and the bottom end of the 95% HPD intervals, respectively. The median of the medians of the posterior estimates is displayed as the darker-colored dot. The true value of the parameter is indicated by the gray horizontal line. In (*A*), (*B*), and (*E*), the *y*-axis has a log-scale. For all subfigures, the summary statistics for the PIQMEE method on the full data set (PIQMEE_all) are shown in red, for the CLASSIC method on the full data set (CLASSIC_all) in blue, for the CLASSIC method on the unique subset (CLASSIC_unique) in green, and for the CLASSIC method on the random subset (CLASSIC_random) in purple.

Results of these analyses show that the estimated death rate is still low, but higher, and closer to the true value, than when the tree is unfixed (fig. 1). However, not all of the bias can be explained by the missing short terminal branches. Thus, the bias in the death rate and R_e, comes partly from the overestimation of the tree height and partly from the missing short terminal branches. Using the CLASSIC_unique method, the R_e, and for most settings also the tree height, are estimated with the lowest accuracy, that is, the RMSE is the highest. Similarly, for each single parameter across all five data set sizes the coverage percentage for the mixed runs (defined as the percentage of analyses, out of those for which an ESS of 200 was reached for all parameters, whose HPD interval includes the true parameter value) is the lowest when the unique sequences only are analyzed (table 1). The coverage ranges between 82% and 99% for R_e and is close to 0% for the death rate. The higher coverage for the R_e can be partially explained by very low precision, that is, very wide 95% HPD intervals, for that parameter.

In contrast to the CLASSIC_unique results, the analyses of random subsets of the data do not lead to bias in the estimates of the population dynamic parameters. For both the R_e and the death rate, the accuracy and the precision are nevertheless lower than for the CLASSIC or PIQMEE analyses of full data sets. Due to the reduced sample size, larger HPD intervals without biases are expected when analyzing random samples as opposed to the full data set. High coverage is therefore expected as a direct consequence of the increased HPD intervals. However, the tree height inferred is generally smaller than

Table 1. Coverage, Accuracy, and Precision of the PIQMEE versus the CLASSIC Method.

Number of Simulated Sequences	Method	Median Number of Analyzed Sequences	Runs Reaching 200 ESS	CPU Seconds per 200 ESS	Tree Height					R_e					Death Rate				
					True	MoM	RMSE	Coverage	HPD Size	True	MoM	RMSE	Coverage	HPD Size	True	MoM	RMSE	Coverage	HPD Size
300	PIQMEE_all	300	100	510.716	189.414	164.452	5.089	92	1.286	1	0.998	0.008	100	0.023	1	1.023	0.352	99	1.352
	CLASSIC_all	300	98	868.493	189.414	167.685	5.098	90	1.291	1	0.998	0.007	98	0.022	1	1.016	0.352	97	1.355
	CLASSIC_unique	19	100	61.457	189.414	153.707	5.670	81	1.135	1	12.932	17.343	99	643.745	1	0.004	0.935	0	0.204
	CLASSIC_random	19	100	51.851	189.414	147.843	5.410	91	1.309	1	0.998	0.177	100	0.031	1	1.025	1.080	96	3.001
1,200	PIQMEE_all	1,200	98	3,457.728	664.673	631.564	10.484	93	1.437	1	0.999	0.004	98	0.006	1	1.031	0.308	95	1.274
	CLASSIC_all	1,200	93	13,950.919	664.673	679.082	10.661	88	1.354	1	0.999	0.002	93	0.006	1	1.032	0.304	91	1.238
	CLASSIC_unique	19	100	91.485	664.673	671.776	12.251	78	1.126	1	11.717	15.352	90	634.224	1	0.001	0.977	0	0.058
	CLASSIC_random	19	100	94.050	664.673	571.125	12.426	84	1.302	1	0.999	1.419	100	0.017	1	1.134	119.208	96	4.041
2,100	PIQMEE_all	2,100	95	7,532.271	1,195.555	1,214.334	12.626	91	1.520	1	1.000	0.001	95	0.004	1	1.119	0.371	92	1.313
	CLASSIC_all	2,100	83	33,230.011	1,195.555	1,346.233	13.353	79	1.417	1	1.000	0.000	83	0.003	1	1.104	0.363	80	1.266
	CLASSIC_unique	19.5	100	115.094	1,195.555	1,196.450	15.974	80	1.217	1	5.426	10.711	93	581.605	1	0.002	0.975	1	0.070
	CLASSIC_random	19.5	97	44.540	1,195.555	1,053.862	14.153	89	1.488	1	1.000	0.917	97	0.008	1	0.985	1.446	92	3.468
3,000	PIQMEE_all	3,000	88	16,509.389	1,802.415	1,900.862	17.910	86	1.471	1	1.000	0.000	88	0.002	1	1.012	0.333	84	1.190
	CLASSIC_all	3,000	82	68,108.129	1,802.415	1,989.483	18.058	81	1.460	1	1.000	0.000	82	0.002	1	1.010	0.343	79	1.151
	CLASSIC_unique	19	100	133.010	1,802.415	1,879.452	21.912	79	1.234	1	6.364	10.479	92	586.983	1	0.001	0.977	0	0.046
	CLASSIC_random	19	97	52.201	1,802.415	1,391.544	21.435	84	1.440	1	1.000	0.119	97	0.005	1	0.921	2.555	96	3.420
6,000	PIQMEE_all	6,000	79	92,081.925	3,685.930	4,097.601	25.200	75	1.407	1	1.000	0.000	79	0.001	1	1.000	0.457	78	1.086
	CLASSIC_all	6,000	20	352,521.226	3,685.930	4,737.040	21.652	19	1.390	1	1.000	0.000	20	0.001	1	0.957	0.781	19	1.023
	CLASSIC_unique	20	100	181.031	3,685.930	4,006.775	31.893	81	1.290	1	3.683	8.946	82	550.668	1	0.001	0.980	1	0.041
	CLASSIC_random	20	95	44.510	3,685.930	3,012.714	25.474	86	1.421	1	1.000	0.214	95	0.005	1	1.068	63.933	92	3.754

NOTE.—Summary statistics for the analyses of 100 simulations per setting using four different methods to analyze the same data sets. The "true" value of the parameter, that is, the value of the parameter under which the data were simulated, is displayed in the respective column for each parameter. For the tree height parameter, the median of the "true" tree heights is displayed. MoM represents the median of medians of those analyses out of 100 that reached ESS 200 for each parameter. RMSE is the root mean squared error of the medians of those analyses out of 100 that reached ESS 200 for each parameter. RMSE was normalized by the true value of the respective parameter. Coverage displays the number of analyses (out of 100) that reached ESS of 200 for each parameter and for which the true value of a parameter was contained in the 95% HPD interval. HPD size is the median size of the 95% HPD interval of those analyses out of 100 that reached ESS of 200 for each parameter. HPD size was normalized by the true value of the respective parameter. The table also shows the median CPU time of those runs that finished within the runtime limit of 7 days and reached 200 ESS for each parameter.

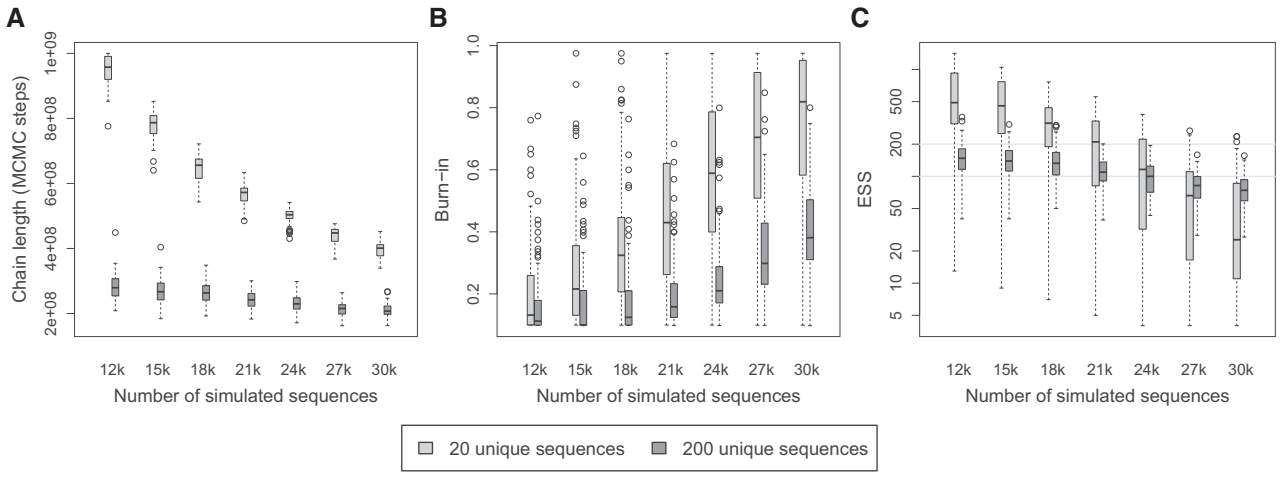

Fig. 2. Performance of the PIQMEE method on very large data sets. We analyzed data sets consisting of 12,000, 15,000, 18,000, 21,000, 24,000, 27,000, and 30,000 sequences with PIQMEE. All sequences were sampled at one point in time (homochronous sampling). (*A*) The chain length reached in 14 days. (*B*) The burn-in, expressed as the fraction of the entire chain, that needed to be removed in order to obtain well-mixing part of the chain. We note that if we have to remove more than half of the chain, it indicates that convergence is not necessarily reached and the chain should thus be run for longer. (*C*) The ESS reached. The *y*-axis in (*C*) has a log-scale.

the true tree height of the full data set. As mentioned in the introduction, this is due to the fact that the random subset of sequences may not contain the two most divergent sequences from the full set, leading to a tree with a younger MRCA. In such a case, the estimate of tree height from the subsample will have seemingly increased bias when compared with the MRCA of the full data set, an observation reproduced in our results (table 1).

The PIQMEE_all and CLASSIC_all analyses lead to very similar parameter estimates, confirming that our PIQMEE implementation works correctly. Both the table 1 and the figure 1 show that the accuracy, precision, and coverage for all parameters inferred with PIQMEE are almost identical to those obtained using the CLASSIC method applied to the full data set. However, the PIQMEE_all analyses are much faster. With increasing number of sequences, the speed difference gets bigger. The reason for this is that the sequences were simulated on each tree such that the median number of unique sequences was around 19–20. This also means that the amount of duplicates for each unique sequence increases with increasing number of total sequences. This in turn results in the speed advantage of the PIQMEE method increasing as the data set grows in size. The analyses with the CLASSIC_all method start to mix very slowly at 2,100 sequences. At 6,000 sequences, many of the CLASSIC_all analyses fail early on due to too large memory requirements. Of those that run many converge slowly, resulting in only 20 out of 100 well-mixed runs. By contrast, when the number of input sequences reaches 6,000, 79 out of 100 PIQMEE runs mix well.

To find out the maximum size of the data set that PIQMEE is able to analyze within a 2-week runtime, we simulated 100 trees for each of following sizes: 12,000, 15,000, 18,000, 21,000, 24,000, 27,000, and 30,000 tips. All tips in the trees were sampled at one point in time. We then simulated sequences on these trees under two different substitution rates, such that there were ~20 and 200 unique sequences. The resulting sequence data sets were analyzed by PIQMEE. Figure 2 and supplementary table S1, Supplementary Material online, show for which data sets we obtained well-mixed MCMC chains.

As expected, with increasing number of sequences in the data set the sampling from the posterior distribution gets slower, as indicated by the decreased number of steps the MCMC chain achieved within the given run time of 14 days. The mixing also gets slower, as indicated by the increasing number of burn-in steps.

The data sets with ~20 unique sequences ran and also converged faster than those with ~200 unique sequences. For the data sets with 20 unique sequences, the majority of the analyses of data sets with 21,000 or less sequences mixed well (ESS$\geq$200). In contrast, none of the data set with ~200 unique sequences mixed well. Irrespective of the number of unique sequences, the majority of the analyses reached ESS of 100 for data sets with 24,000 and less sequences.

Method Performance under Various Scenarios

Sequence Evolution Models

We have shown above that the parameter inference under PIQMEE works very well when the sequences are simulated and analyzed with the JC69 model under the strict molecular clock model. The phylogenetic likelihood calculation under PIQMEE also works for more complex models of sequence evolution. The parameter inference is equally good under PIQMEE as under CLASSIC method with the full data set when the sequences are simulated with HKY (Hasegawa–Kishino–Yano) or GTR (general time reversible) model (see supplementary fig. S2, Supplementary Material online, for an example of data sets with 300 simulated sequences).

Heterochronous Sampling

In addition to the homochronous sequences (sampling at one point in time) shown above, our PIQMEE method can

also handle data sets with heterochronous sequences, that is, sampled through several points in time. Again, as expected given a correct PIQMEE implementation, the parameter estimates under PIQMEE_all and CLASSIC_all are the same (supplementary fig. S3 and table S2, Supplementary Material online), despite an additional parameter, that is, clock rate, being estimated. Similarly to the analysis of homochronous sequences, the analysis of heterochronous sequences by the PIQMEE method is faster than by the CLASSIC method. In fact, the speed advantage of PIQMEE is more pronounced on the heterochronous data set as compared with the homochronous data set. With just 300 sequences being analyzed, PIQMEE is, based on the median runtime, five times faster on the heterochronous data set (3,203 vs. 16,208 CPU seconds), whereas it is 1.7 times faster on the homochronous data set (511 vs. 868 CPU seconds).

Furthermore, we obtain the same patterns with respect to speed, bias, accuracy, and precision for the analyses of the unique subsets of the data as we did in the homochronous case. The only exception is that in the homochronous scenario the R_e was overestimated, whereas in the heterochronous data sets it is underestimated.

For the random subsets, the results are generally less accurate than in the homochronous scenario. The CLASSIC_random method overestimates the clock rate and the death rate and underestimates the R_e. This is most likely due to a lack of information content in the subsampled data set, insufficient for the inference of all parameters of the model. When we reanalyze the same data set with a stronger prior around the true value of the death rate parameter (supplementary fig. S4, Supplementary Material online), estimates of the death rate and clock rate improve.

Clock Models

Analyses of subsets of the (heterochronous) data using the relaxed clock model lead to similar results to those observed when the strict clock model is used. Both the CLASSIC_random and CLASSIC_unique methods are still faster than full data set analyses; however, they also lead to biased parameter estimates. We further observe that there is a slight deviation of the PIQMEE method results from those of the CLASSIC method when applied to full data sets (supplementary fig. S5, Supplementary Material online). The main difference is that the confidence intervals for the R_e and the average clock rate parameter are larger for the PIQMEE method as compared with the CLASSIC method. The coverage of the PIQMEE method is very good, and similar to that of the CLASSIC_all method, for all parameters displayed in supplementary figure S5, Supplementary Material online, with exception of the death rate (see supplementary tables S3 and S4, Supplementary Material online). In addition, the PIQMEE method tends to slightly underestimate the death rate and the clock rate, while overestimating the tree height. This pattern is seen even if the tree is fixed (supplementary fig. S6, Supplementary Material online) to the true tree. The distribution of substitution rates associated with different types of branches (internal vs. external) in our tree (supplementary figs. S7 and S8, Supplementary Material online) makes it clear that the bias in the parameter estimates is a result of the PIQMEE method's assumption that each haplotype only evolves once within the tree. This assumption translates into the PIQMEE method requiring that for each haplotype subtree the sequence at the MRCA of the subtree is exactly the same as the sequence at the tips. This has two consequences that are reflected in our results. Firstly, it can be seen from the transition probability formula (see the Materials and Methods section, phylogenetic likelihood) that the lower the substitution rate for a fixed, large haplotype subtree is, the higher the phylogenetic likelihood value will be. This association between low rates and large subtrees can be seen in our results (supplementary figs. S7 and S8, Supplementary Material online). In an unfixed tree, this can then translate in forcing the root of the tree to be older, and by consequence the death rate (as well as the birth rate) to be lower.

Secondly, the PIQMEE method forces the sequence within haplotype subtrees to remain unchanged until further in the past than the CLASSIC method does. The (internal) branches in the tree that are above the haplotype subtrees need to take relatively higher substitution rates (see supplementary figs. S7 and S8, Supplementary Material online), such that the sequences change fast enough between the MRCA of the haplotype subtrees and the next internal node, where they join with another tip or MRCA of another haplotype's subtree.

Skyline Model

Finally, the tree prior in the PIQMEE method has been implemented as an extension of a BDsky model. PIQMEE preserves the skyline functionality of BDsky and is thus able to capture the changes in the population dynamics parameters (R_e and death rate) over time (see supplementary fig. S9, Supplementary Material online). As in the previous analyses, where the sequences were sampled at multiple time points, both the CLASSIC_random and the CLASSIC_unique method perform poorly. In contrast to the nonskyline methods, the clock rate is severely underestimated by both CLASSIC_random and CLASSIC_unique skyline methods. This leads to overestimation of the tree height. Also the population dynamics parameter estimates have large biases, with both methods overestimating the R_e and underestimating the death rate. These biased parameter estimates are very likely due to the lack of sufficient information in the data for inference of all parameters of the model, confirmed by the results of the analyses with more defined priors (see supplementary fig. S10, Supplementary Material online).

Analysis of the Real Within-Host HIV Data Sets

We applied the PIQMEE method to the publicly available HIV sequence data from 8 out of 11 patients published in Zanini et al. (2015). Two patients, patient 4 and patient 7, were omitted because they have been superinfected. One additional patient, patient 10, was removed from the analyses, because for the genomic region we analyzed, sequences were successfully obtained for one time point only,

Table 2. Summary of the HIV Data Set.

Patient Number	Estimated Time since Infection (years)	Total Number of Sequences	Number of Unique Sequences
p1	8.21	10,394	52
p2	5.53	6,998	13
p3	8.44	3,625	41
p5	5.89	7,408	35
p6	7.00	10,798	15
p8	4.96	4,689	31
p9	8.10	5,253	17
p11	5.60	4,693	35

NOTE.—The second column shows clinically established estimated time since infection for each patient. The third column shows the total sequence count covering the C2-V5 region of HIV genome. The fourth column shows the number of unique sequences in each patient data set.

insufficient to perform inference without fixing the clock rate. Table 2 shows summary statistics of the analyzed data sets.

We used a skyline model with 1, 3, or 5 intervals for R_e. We ran the analyses once with and once without sequences (option "Sample from prior" in BEAST 2) to check how the priors interfere with each other and whether the distributions obtained from the two runs are different, that is, whether the sequence data contain enough information to allow for inference of the parameters of interest. Running the analyses when excluding sequences in BEAST 2 is not equivalent to running the analyses under the prior in the Bayesian sense, as the birth–death model still uses the sampling dates of the sequences as source of information (Boskova et al. 2018).

None of the runs using the CLASSIC_all method mixed within the 7-day runtime. We obtained well-mixed chains for both when using the sequence data and when excluding the sequence data from the PIQMEE analyses for the following patients and settings: patients 2, 8, and 9 when using one interval for R_e, patient 8 when using three intervals for R_e, and patient 5 when using five intervals for R_e (see supplementary table S5, Supplementary Material online). However, in the analyses of patients 2 and 9, when using a single interval for R_e in the PIQMEE method, the distribution of the R_e parameter was essentially invariant to the inclusion or exclusion of the sequence data. This indicates that the sequence data bring very little information to the model on this parameter (see supplementary figs. S11 and S12, Supplementary Material online, for patients 2 and 9, respectively).

In addition to the analyses of full data sets with the CLASSIC and PIQMEE methods, we also analyzed the data set consisting of only the unique sequences (see table 2 for unique sequence counts) and a random subset of 600 sequences. For patients 2 and 8 when using one interval for R_e (supplementary figs. S11 and S13, Supplementary Material online) and for patient 8 when using three intervals for R_e (supplementary fig. S14, Supplementary Material online), only the analyses with PIQMEE_all and CLASSIC_unique methods mixed well. There was only a single case for which all three analyses (PIQMEE_all, CLASSIC_random, and CLASSIC_unique) mixed well both with and without the sequence data: when using one interval for R_e with patient 9 data (supplementary fig. S12, Supplementary Material online). However, same as for PIQMEE_all, inclusion or exclusion of the sequence data in the CLASSIC_random analyses did not make a difference for the distribution of R_e. Finally, only the analyses with PIQMEE_all and CLASSIC_random mixed well when using five intervals for R_e for patient 5 data (fig. 3).

Across the successfully converged analyses (fig. 3 and supplementary figs. S11–S14, Supplementary Material online), the clock rate is estimated to be higher by the CLASSIC_random and CLASSIC_unique methods than the PIQMEE_all method. Additionally, the tree height and the R_e estimates provided by the CLASSIC methods often differ from those obtained with PIQMEE_all method. When using one interval for R_e, the R_e estimates by the CLASSIC methods are always below the PIQMEE_all estimates. In addition, for patient number 5 and five intervals for R_e (fig. 3), the PIQMEE method estimates that the HIV population was relatively stable (R_e is around or slightly above 1) since the time of infection. The virus population was growing the fastest between 1 and 2.3 years before the last sample, and this growth continued, though at a slower pace in the year immediately preceding the last sample. However, the random data set does not capture this trend well. The CLASSIC_random method estimates that the virus population was increasing ($R_e > 1$) between 1 and 2.3 years before the last sample and decreased ($R_e < 1$) in the year immediately preceding the last sample.

In summary, the analyses of the random and unique subsets of real HIV data sets provide very different parameter estimates as compared with the analyses of full data sets using the PIQMEE method, which is consistent with the analyses of the simulated data sets.

Discussion

Understanding the dynamics of pathogen dissemination is crucial for introducing appropriate measures to either stop or slow down its spread. There are two different scales at which information about pathogen dynamics can be gained: between- and within-host. For most infectious diseases, the between-host level is of importance. This is especially the case for infections that spread quickly among individuals, for example, the recent Ebola epidemic (Althaus 2014), the Zika virus epidemic (Ferguson et al. 2016), or the ongoing Covid-19 pandemic (Ferguson et al. 2020). If the infection is long-lasting with damaging effects to the host, such as HIV or HCV, insights into within-host disease progression and pathogen evolution are necessary for effectively personalizing treatment

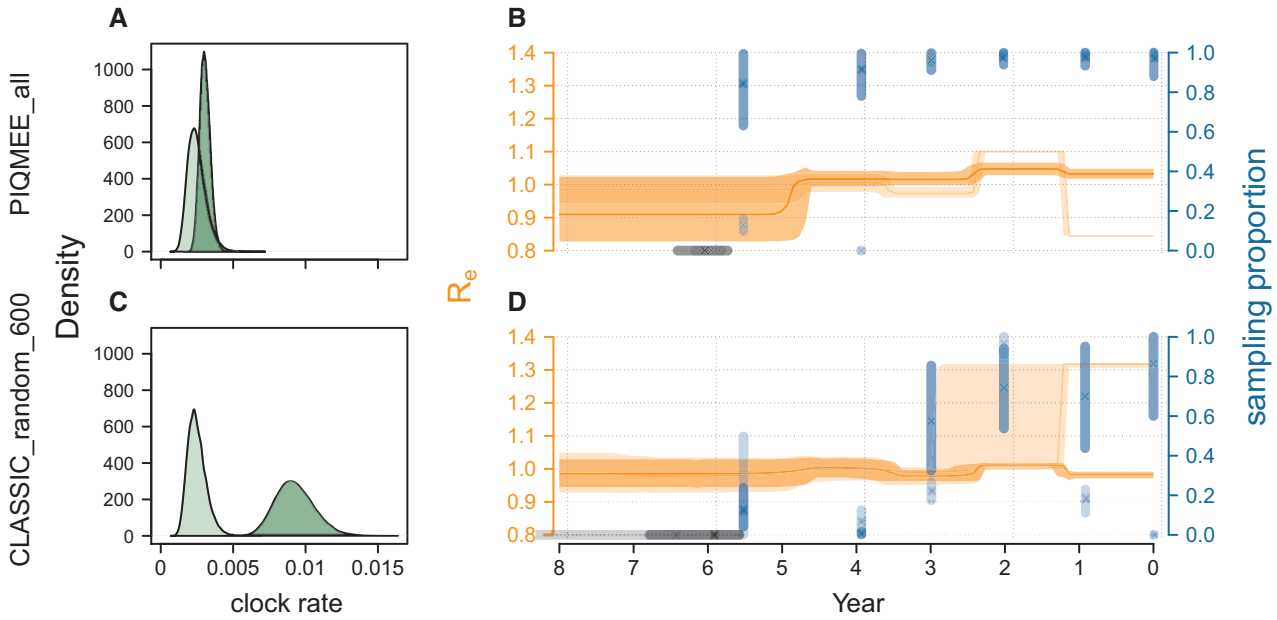

FIG. 3. Skyline plot of chronically infected HIV patient number 5 from Zanini et al. (2015), analyzed using the skyline model with five intervals for R_e. We plot the distribution of the clock rate (*A* and *C*), and the tree height, the effective reproductive number R_e (orange) as well as the sampling proportion (blue) (*B* and *D*). (*A*) and (*B*) correspond to analyses performed with the PIQMEE_all method. (*C*) and (*D*) The results of analyses with the CLASSIC_random method. The results of the CLASSIC_unique method are not displayed here because the analyses when the sequence data were excluded did not mix well. For all analyses, the death rate was fixed to 124 per year. The distributions obtained when excluding the sequence data are shown in lighter color, whereas the distributions obtained when the sequence data are included are plotted in darker shades. The gray interval shows the distribution of the tree height. Note that when excluding the sequence data, the distribution of tree height for the random subsample ranged from 5.6 to 14.6 years (the plot is truncated at 8 years). The clock rate in plots (*A*) and (*C*) is in units of substitutions/site/year. The time on the *x*-axes of plots (*B*) and (*D*) goes from the time of the last sample (0) backwards and is displayed in units of calendar years.

of individual patients (Wei et al. 1995; Perelson et al. 1996; Gray et al. 2011; Ribeiro et al. 2012).

The within-host populations are often sequenced using next-generation sequencing methods. Advances in sequencing technology provide us with an in-depth view into pathogen population diversity (Schuster 2008; Zagordi et al. 2011; Töpfer et al. 2013; Pandit and de Boer 2014). Many studies report that these within-host data sets contain many biological duplicates (Boeras et al. 2011; Töpfer et al. 2014; Wu et al. 2014). The development of sophisticated phylodynamic methods to fully exploit such data is now apt.

We have shown through simulations that analyses using only the unique sequences lead to biased parameter estimates. Additionally, when compared with analyses of full data set, analyses using a random subset of the data show decreased precision. Analyses of random subsets of data can lead to parameter estimates, for example, of the tree height, that do not correctly reflect the properties of the full data set. Furthermore, the analyses of full data sets with the classic method slowed down significantly with increasing number of sequences. In fact, only 20% of the analyses with 6,000 sequences reached ESS of 200 for all parameters, despite the fact that the complexity of the data set, as measured by the number of unique sequences, remained the same across all data sets.

In current Bayesian phylodynamic methods, each sequence corresponds to a separate tip in a tree. The inclusion of duplicate sequences means that tree space increases significantly but not in an informative way. Identical sequences can be freely exchanged on the tree without the phylogenetic likelihood and the tree prior changing. Such inefficient tree space exploration causes poor mixing of MCMC chains. Long chains are therefore needed to achieve satisfactory ESSs for all estimated parameters. This is especially problematic for, and may thus completely preclude, analysis of large data sets with many duplicate sequences.

We have therefore proposed a new method, PIQMEE, that takes advantage of the fact that duplicate sequences can be treated differently than separate tips in the tree, and thus only unique sequences are represented as tips in the tree. It is founded on the observation that the topology of duplicate sequences cannot be resolved, but the timing of their branching can, because this is informed by a combination of the phylogenetic likelihood $f[\mathrm{D}|\tau,\theta]$ and the tree prior $f[\tau|\eta]$. PIQMEE considers trees on unique sequences only, keeping track of the branching times of duplicates. This tree space can be efficiently explored with current (adapted) and additional new BEAST 2 mechanisms that propose new tree states (operators). The main assumption of the method is that each haplotype arises only once through mutation.

We have shown that the PIQMEE method is as accurate and as precise as the classic implementation of the likelihood and tree prior for the full data set. The PIQMEE method is capable of analyzing both homochronous and heterochronous samples. The estimates of parameters are the same as in analyses using the classic method for data sets evolving under the strict clock model.

It is also possible to use the PIQMEE method with relaxed clock models of sequence evolution. However, slight biases in the clock rate, the tree height, and the death rate may arise due to the PIQMEE assumption of each unique sequence arising only once within the tree. The inference will thus show slight deviations from the classic method results.

We have shown that the PIQMEE method tied with the strict clock model can process data sets of up to 21,000 sequences, of which 20 were unique and the rest were duplicates. This means that larger amounts of data can now be processed using Bayesian phylodynamic methods than was possible before. By being able to use all available sequences and not only the unique subset of them, the MRCA (proxy for the start of the infection) and the within-host pathogen population dynamics can be studied in more detail.

The composition of quasispecies populations, in terms of their unique sequence spectrum and frequency, is temporally dynamic (Bull et al. 2011; Pandit and de Boer 2014). This can be attributed to factors such as sequence adaptation to host immune system and population bottlenecks due to the patient's drug regimen (Pybus and Rambaut 2009). Detailed information on the pathogen population composition and dynamics will yield insight into the dependency (of speed) of drug resistance development on population composition and history. The bottlenecks could be identified and correlated with events such as a change of the drug regimen. Similarly, if a sudden expansion in diversity of quasispecies was seen, this could be correlated with events such as a failure of the patient to stick to the treatment. The PIQMEE method is ideal for the study of large data sets with many duplicate sequences, for example, those obtained from chronic infections of a host with pathogen such as HIV or HCV. We have shown the usefulness of our method for such data sets by successfully applying it to large sequence data sets from patients chronically infected with HIV.

Within-host population dynamics is an important factor to include in the model if one wants to correctly reconstruct the between-host transmission network (Didelot et al. 2014; Romero-Severson et al. 2016; Didelot et al. 2017). Although for some studies using a single sequence from a rich within-host quasispecies seemed to be sufficient to approximate the date of infection (Poon et al. 2011), others claim that one sequence per patient may not be enough to allow for correct reconstruction of the transmission chain (Ypma et al. 2013; Worby et al. 2014; Volz et al. 2017). Corrections in the form of modeling the within-host population dynamics are necessary (Ypma et al. 2013; Didelot et al. 2014, 2017; Klinkenberg et al. 2017; Volz et al. 2017) even if several unique sequences are used (Vrancken et al. 2014). Only the correct transmission network can lead to reliable parameter estimates of the transmission dynamics (Ypma et al. 2013; Volz et al. 2017). Using a model that can accommodate many sequences from a single host while reconstructing the between-host transmission network should lead to more reliable estimates of the transmission network structure and dynamics. Our method can serve as a starting point for designing such models.

Several of the nested within- and between-host models treat each patient as a separate compartment. Often, the within-host dynamics are modeled using a coalescent approach, with coalescent events among sequences being allowed to happen only within, but not between these compartments (Hall et al. 2015; De Maio et al. 2016). For the PIQMEE method to be compatible with these approaches, the tree prior would need to be adapted to the coalescent framework. For a coalescent event between two identical sequences, the coalescent rate would depend not only on the number of identical sequences present in that host at that time point (see calculation of the factor γ for the tree prior in the Materials and Methods section) but also on the population size of the pathogen within that host.

Although the PIQMEE method has originally been conceived for quantifying within-host quasispecies evolution and dynamics, there is no reason why the method could not be applied to other data sets, such as between-host data sets, where many sequences are identical. The only difference to the within-host model would be the meaning of the parameters, for example, the R_e in the within-host context would represent the ability of the pathogen to spread within the host, whereas for the between-host dynamics the R_e would refer to the between-host spread.

A drawback of our PIQMEE method is that it requires nonrecombining sequences as input. Although recombination occurs in many pathogens (Simon-Loriere and Holmes 2011), only a few phylodynamic methods can currently handle such data sets (Bloomquist and Suchard 2010; Vaughan et al. 2017). Using portions of the genes or genomes that are known or assumed to not be recombining (González-Candelas et al. 2011; Smyth et al. 2014) is the usual workaround to this problem, and we would recommend this approach if using PIQMEE with such data sets.

In summary, the PIQMEE method is a significant step toward faster analysis and accurate estimation of population dynamics based on deeply sequenced quasispecies data sets, or any other large data set with a high proportion of duplicate sequences. The method could further be improved by implementing the phylogenetic likelihood calculation to work with BEAGLE (Suchard and Rambaut 2009; Ayres et al. 2019), by making corrections such that the relaxed clock models would be fully compatible with the PIQMEE model assumptions, and by allowing for recombination of sequences. Also, tree priors other than the birth–death-based BDsky model could be implemented to work with our tree structure allowing for analyses under various population dynamic models.

Materials and Methods

PIQMEE Method Description

New Tree Structure

As mentioned in the introduction, PIQMEE uses a reduced tree structure (fig. 4). It is composed of the tree of unique sequences, τ_u, an array of branching times (colored dashes on the branches of τ_u in fig. 4), and sampling times of all the duplicate sequences, denoted ζ. Note that each unique sequence (haplotype) corresponds to exactly one tip in the tree, and that tip represents the most recently sampled duplicate of the haplotype.

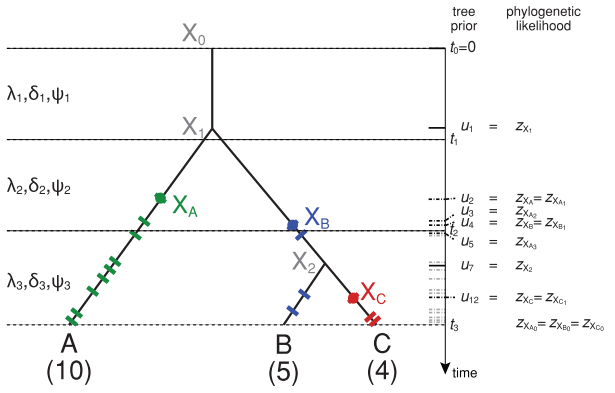

Fig. 4. Reduced representation of a tree of three unique haplotypes (A–C) and their duplicates. The origin of the process (the tree) is denoted as X_0. We call the internal nodes that join the unique haplotypes (X_1, X_2) the bifurcation nodes of the tree. The total count for each haplotype is noted below the corresponding tip in the tree. Note that the branching times of the duplicates are depicted as colored dashes on the tree. The notation of these times in the tree prior and the phylogenetic likelihood is shown to the right. We only show the times of the bifurcation nodes (solid lines leading to the time axis), and first few branching times of duplicates (dark dash-dotted lines). Due to space constraints, the rest of the branching times of duplicates are only shown as gray dash-dotted lines. The last sampling time of each haplotype is shown (z_{A_0}, z_{B_0}, z_{C_0}). In our example, these times are the same for all three haplotypes. The colored dots on the tree (X_A, X_B, X_C) represent the start of a haplotype and coincide with the first branching event of haplotype (A, B, C, respectively). From this point, all the way to the tip we assume the sequence is not mutating (see the Materials and Methods section on Phylogenetic model). Note that the haplotypes can start above any bifurcation node on a path between the tip corresponding to that haplotype and the origin of the tree. If a bifurcation node is on the path between the haplotype's first branching point and the tip, we say that the haplotype passes through that node, for example, the haplotype B is passing through bifurcation node X_2. This in particular means that we assume that the sequence at node X_2 is identical to the sequence at tip B. We use birth–death skyline model for the analyses. Here, the process is split into three intervals, delimited by times t_0, t_1, t_2, t_3 (black dashed lines). Population dynamic parameters (λ, δ, ψ) are estimated for each interval separately.

Let the number of unique sequences be n_u. The letters $X_i, i \in \{1, \ldots, n_u - 1\}$, denote the internal nodes in the tree τ_u and the capital letters A, B, ... denote the tips. For the example tree in figure 4, τ_u is a tree with internal (bifurcation) nodes X_1 and X_2 and three tips labeled A, B, C, which represent the haplotypes A, B, and C, respectively.

The total count of copies of a given haplotype H is denoted as N_H. Each subtree for N_H copies of the haplotype H sequences thus has $N_H - 1$ times where the duplicate sequences branch off. The duplicate sequences branch off at degree-2 nodes, that is, at the nodes connecting two edges, or branches in τ_u. We denote these degree-2 nodes X_{H_i}, where H_i refers to the ith duplicate of H. Note that the index 1 points to the time of the top-most (first) branching event of the duplicate of a given haplotype. For example, X_{A_1} is the branching point of the first duplicate of haplotype A, depicted as the top-most dash of haplotype A in figure 4. X_{A_1} coincides with X_A, the "start" of the haplotype A and is marked as a dot on the tree in figure 4. X_{A_2} is the branching point of the second duplicate, arising just below the first one (second dash from the top), etc.

Phylogenetic Model

We denote the time at which the bifurcation node X_i occurs in the tree as z_{X_i}. The branching time of the ith duplicate of haplotype H is denoted as $z_{X_{H_i}}$, and the sampling time of that duplicate sequence is denoted as z_{H_i}. Note that the sampling time of the haplotype representative sequence (the sequence in τ_u) is z_{H_0} and corresponds to the latest sampling time of H. Let the time flow forward, that is, starting at time 0, some time in the past, with a single individual, then we have the sorted times arrays such that $z_{X_1} \leq z_{X_2} \ldots \leq z_{X_{nu-1}}$, $z_{H_1} \leq z_{H_2} \leq \ldots \leq z_{H_{N_H}}$, and $z_{X_{H_1}} \leq z_{X_{H_2}} \leq \ldots \leq z_{X_{H_{N_H-1}}}$. The edge length between two nodes, for example, between X_i and X_j is the absolute value of the distance in time of the two nodes and is denoted as $|z_{X_i} - z_{X_j}|$.

The lower case letters x_i or $a, b, \ldots$ denote the corresponding sequences at the nodes X_i or A, B, ..., respectively. We assume that the sequences at the tips are known and are exactly the sequences of the corresponding haplotype H, which is denoted as h. For the example shown in figure 4, the tips A, B, and C represent sequences a, b, and c, respectively. We denote by x_i the sequence at the bifurcation node X_i, where, for example, $x_1 = G$ means that the sequence at node X_1 is one single nucleotide, namely G.

The substitution model parameters are composed of $\theta = \{Q, \Pi\}$, where Q is the substitution rate matrix defining the transition probabilities P and Π is the vector of stationary distribution for each state in the state space S. When dealing with nucleotide sequences, $S = \{A, C, G, T\}$.

As our tree τ_u only represents the topology of the unique sequences, we needed to rewrite the phylogenetic likelihood formulated using the Felsenstein's peeling algorithm (Felsenstein 1981) to fit such a structure. Instead of the full tree, τ, the adapted Felsenstein likelihood takes τ_u and ζ as input.

The new phylogenetic likelihood formula accommodates our main assumption of each haplotype arising only once during the history of the process (i.e., in the tree) by imposing that the duplicates of a haplotype do not mutate at all. Thus, the probability of no mutation event on a branch with length t is $\tilde{P}_{n(i)\to n(i)}(t) = e^{q_{ii}t}$ with $q_{ii} = -\sum_{j=1}^{|S|} q_{ij}$ $(i, j = \{1, \ldots, |S|\}; n(i) \in S)$.

There are two direct consequences of our method's assumption reflected in the formula. First, the haplotype sequence stays the same for the duration of the sum of the branch lengths defined by the distances from each of the duplicate sequence branching points X_{H_i} (dashes in fig. 4) to the corresponding sampling time z_{H_i}. For a single site, and N copies of a haplotype, we can write, $\prod_{j=1}^{N} \tilde{P}_{n(i)\to n(i)}(|z_{X_{H_j}} - z_{H_j}|) = \tilde{P}_{n(i)\to n(i)}(\sum_{j=1}^{N} |z_{X_{H_j}} - z_{H_j}|)$. Second, the sequence at each degree-2 node X_{H_i} is known, such as $x_{H_1} = x_{H_2} = \cdots = h$. Thus, the sequence at each degree-2 node belonging to haplotype H is h, the same as the sequence

at the tip H. Again, for our example shown in figure 4, $x_{B_1} = x_{B_2} = x_{B_3} = x_{B_4} = b$. In addition, if a haplotype passes a bifurcation node, that node can only have a sequence identical to the sequence corresponding to the passing haplotype. In figure 4, haplotype B passes through the bifurcation node X_2; therefore, the node X_2 can only have a sequence identical to the sequence B, that is, $x_2 = b$. However, for the bifurcation nodes through which no haplotype passes, the exact sequence is unknown and we thus need to integrate over all possible states it can assume, meaning that in the example in figure 4, for node X_1, we sum over all states in S at every position of the sequence.

Based on this, and for simplicity assuming the sequence alignment only contains a single site, the phylogenetic likelihood can be written as:

$$f[\mathrm{D}|\tau_u,\zeta,\theta] = \sum_{x_1\in S}\Pi_{x_1}\left[P_{x_1\to a}\left(\left|z_{X_1}-z_{X_A}\right|\right)\tilde{P}_{a\to a}\left(\left|z_{X_A}-z_{A_0}\right|\right)\tilde{P}_{a\to a}\left(\sum_{j=1}^{N_A-1}\left|z_{X_{A_j}}-z_{A_j}\right|\right)\right]$$
$$\left[P_{x_1\to b}\left(\left|z_{X_1}-z_{X_B}\right|\right)\tilde{P}_{b\to b}\left(\left|z_{X_B}-z_{X_2}\right|\right)\left[\tilde{P}_{b\to b}\left(\left|z_{X_2}-z_{B_0}\right|\right)\tilde{P}_{b\to b}\left(\sum_{k=1}^{N_B-1}\left|z_{X_{B_k}}-z_{B_k}\right|\right)\right]\right.$$
$$\left.\left[P_{b\to c}\left(\left|z_{X_2}-z_{X_C}\right|\right)\tilde{P}_{c\to c}\left(\left|z_{X_C}-z_{C_0}\right|\right)\tilde{P}_{c\to c}\left(\sum_{l=1}^{N_C-1}\left|z_{X_{C_l}}-z_{C_l}\right|\right)\right]\right].$$

The derivation of this formula is equivalent to the derivation of the Felsenstein pruning likelihood. The parts highlighted in blue show where our formula is identical to the classic phylogenetic likelihood. Further, setting $z_{X_H} = z_{X_{H_0}}$, and using the fact that entries of $\tilde{P}$ are exponential functions, we get:

$f[\mathrm{D}|\tau_u,\zeta,\theta] =$

$$\sum_{x_1\in S}\Pi_{x_1}P_{x_1\to a}(|z_{X_1}-z_{X_A}|)\tilde{P}_{a\to a}(\sum_{j=0}^{N_A-1}|z_{X_{A_j}}-z_{A_j}|)$$
$$P_{x_1\to b}(|z_{X_1}-z_{X_B}|)\tilde{P}_{b\to b}(\sum_{k=0}^{N_B-1}|z_{X_{B_k}}-z_{B_k}|)$$
$$P_{b\to c}(|z_{X_2}-z_{X_C}|)\tilde{P}_{c\to c}(\sum_{l=0}^{N_C-1}|z_{X_{C_l}}-z_{C_l}|).$$

For the Jukes–Cantor substitution model (Jukes and Cantor 1969), where all the rates in the substitution rate matrix are the same, say μ, we can rewrite the above equation as follows:

$f[\mathrm{D}|\tau_u,\zeta,\theta=\{\mu,\Pi\}] =$

$$\sum_{x_1\in S}\Pi_{x_1}P_{x_1\to a}(|z_{X_1}-z_{X_A}|)\exp(-\frac{3}{4}\mu\sum_{j=0}^{N_A-1}|z_{X_{A_j}}-z_{A_j}|)$$
$$P_{x_1\to b}(|z_{X_1}-z_{X_B}|)\exp(-\frac{3}{4}\mu\sum_{k=0}^{N_B-1}|z_{X_{B_k}}-z_{B_k}|)$$
$$P_{b\to c}(|z_{X_2}-z_{X_C}|)\exp(-\frac{3}{4}\mu\sum_{l=0}^{N_C-1}|z_{X_{C_l}}-z_{C_l}|).$$

Tree Prior Model

The tree prior within PIQMEE is an extension of the BDsky model (Stadler et al. 2013). Let λ be the birth rate, δ the death rate (referred to as "total rate of becoming noninfectious" in Stadler et al. [2013]), ψ the sampling rate through time, ρ the sampling probability at the time of special sampling effort, $T = (t_1,\ldots,t_m)$ the vector of m times where the rates shift, and S the probability that there is at least one sampled individual at the last sampling time point. Some sequences are sampled during special sampling efforts and the times associated with such efforts are referred to as ρ-times. Sequences sampled outside of these ρ-times are referred to as sequentially sampled and are sampled with rate ψ. We consider trees with N_i sequences sampled at each t_i time, such that $\sum_{i=1}^{m}N_i = N$, and n sequences sampled sequentially. Let us denote by n_i the number of lineages present at the time t_i. These lineage counts n_i can be calculated from τ_u and ζ. We assume that the process starts at time $t_0 = 0$, also called origin, with a single individual. We define the union of z_{X_i} and $z_{X_{H_i}}$ (both introduced in the Phylogenetic Model section above) as $\{u_1,u_2,\ldots,u_{N+n-1}\}$, such that $u_1 \le u_2 \le \ldots \le u_{N+n-1}$. In addition, let $\{v_1,v_2,\ldots,v_n\}$, such that $v_1 \le v_2 \le \ldots \le v_n$ be the times of the sequences that were sampled sequentially.

As the reduced tree structure τ_u and ζ keep track of the duplicates' branching times but not of the full duplicate sequence tree topology, a single reduced tree may represent many different full topologies. In order to account for this, we multiply the BDsky tree prior function of τ_u by a scaling factor γ. Let γ_i be the number of possible ways the lineages can merge at the ith bifurcation or degree-2 node. When the ith node is a bifurcation node, γ_i is simply the number of haplotype H duplicate sequences to which the node can be attached at the time u_i. The exact identity of the haplotype H for each bifurcation node is always defined by the structure of τ_u. When the ith node is a degree-2 node, γ_i equals to $\binom{k}{2}$ and represents all the possible ways the k duplicates of the corresponding haplotype H that exist after the time u_i can merge to $k-1$ duplicates (see supplementary fig. S15, Supplementary Material online, for a visual explanation). Thus, γ_i is defined by the tree structure τ_u as well as the times array ζ, and we always have $\gamma_i \ge 1$. The tree prior $f[\tau_u,\zeta|\lambda,\delta,\psi,\rho,T,S]$ is then obtained based on Stadler et al. (2013),

equation (5) in Supplementary Information, by multiplying the branching events with the corresponding γ_i values:

$$f[\tau_u, \zeta|\lambda, \delta, \psi, \rho, T, S] = \frac{q_1(0)}{1-p_1(0)} \prod_{i=1}^{N+n-1} \gamma_i \lambda_{l(u_i)} q_{l(u_i)}(u_i) \prod_{i=1}^{n} \frac{\psi_{l(v_i)}}{q_{l(v_i)}(v_i)} \prod_{i=1}^{m} \rho_i^{N_i} (1-\rho_i)^{n_i} q_{i+1}(t_i)^{n_i}.$$

The factors $p_i(t)$ and $q_i(t)$ are defined as in Stadler et al. (2013), equations (1) and (3) in the Supplementary Information. We denote by $q_i(t)$ the probability density of an individual at time t giving rise to an edge in an interval of time delimited by t and t_i. In addition, $p_i(t)$ is the probability that an individual alive at time t, where $t_{i-1} \leq t < t_i$, for $i = 1, \ldots, m$, does not leave any sampled descendants at the end of the process. Finally, $l(t)$ is defined as $l(t) = i$ for $t_{i-1} \leq t < t_i$, $i = 1, \ldots, m$.

Implementation in BEAST 2 Software

In order to implement the PIQMEE method in BEAST 2, we rewrote the basic tree class of BEAST 2 to accommodate the new tree structure. If the duplicate sequences of the same haplotype are sampled at different time points, all such time points are merged to a single representative tip to fulfill the assumption that each sequence arose only once during the process. The tip representing each haplotype in τ_u is assigned the date corresponding to the most recent sampling time of that haplotype sequence. The different sampling times, counts of the duplicates, and their branching times are tracked internally. The BDsky model (Stadler et al. 2013) was adapted for the phylodynamic inference, as was the general implementation of the phylogenetic likelihood function $f[D|\tau, \theta]$ for the phylogenetic inference. Modifications to these functions are described above. Lastly, current tree operators were adapted and new tree operators were designed, to allow the MCMC to efficiently explore the new tree space. For adaptation of operators and details on implementation, see the Supplementary Material online.

Simulations and (Real Data) Analyses

For details on how we simulated the sequences, performed the analyses of simulated and real data sets, please refer to the Supplementary Material online.

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Acknowledgments

We would like to thank to Joëlle Barido-Sottani, Denise Kühnert, Timothy G. Vaughan, Alexei J. Drummond, and Stephen Crotty for helpful discussions. We thank Stephen Crotty and Chi Zhang for critically reading parts of the manuscript. We also would like to thank two anonymous reviewers for valuable comments. This work was supported by the ETH Zurich and European Research Council under the Seventh Framework Programme of the European Commission (PhyPD: Grant Agreement No. 335529). V.B. would additionally like to thank Swiss National Science foundation for funding (Early PostDoc.Mobility Fellowship grant number: P2EZP3_184543).

Author Contributions

V.B. and T.S. designed the method and the analyses. V.B. implemented and tested the method within the BEAST 2 framework. V.B. performed the analyses. V.B. and T.S. interpreted the results and wrote the manuscript.

References

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

Althaus CL. 2014. Estimating the reproduction number of Ebola virus (EBOV) during the 2014 outbreak in West Africa. *PLoS Curr*. doi: 10.1371/currents.outbreaks.91afb5e0f279e7f29e7056095255b288.

Ayres DL, Cummings MP, Baele G, Darling AE, Lewis PO, Swofford DL, Huelsenbeck JP, Lemey P, Rambaut A, Suchard MA. 2019. BEAGLE 3: improved performance, scaling, and usability for a high-performance computing library for statistical phylogenetics. *Syst Biol*. 68(6):1052–1061.

Beerenwinkel N, Günthard HF, Roth V, Metzner KJ. 2012. Challenges and opportunities in estimating viral genetic diversity from next-generation sequencing data. *Front Microbiol*. 3:329.

Bloomquist EW, Suchard MA. 2010. Unifying vertical and nonvertical evolution: a stochastic ARG-based framework. *Syst Biol*. 59(1):27–41.

Boeras DI, Hraber PT, Hurlston M, Evans-Strickfaden T, Bhattacharya T, Giorgi EE, Mulenga J, Karita E, Korber BT, Allen S, et al. 2011. Role of donor genital tract HIV-1 diversity in the transmission bottleneck. *Proc Natl Acad Sci U S A*. 108(46):E1156–E1163.

Boskova V, Stadler T, Magnus C. 2018. The influence of phylodynamic model specifications on parameter estimates of the Zika virus epidemic. *Virus Evol*. 4(1):vex044.

Bouchard-Côté A, Sankararaman S, Jordan MI. 2012. Phylogenetic inference via sequential Monte Carlo. *Syst Biol*. 61(4):579–593.

Bouckaert R, Heled J, Kühnert D, Vaughan T, Wu CH, Xie D, Suchard MA, Rambaut A, Drummond AJ. 2014. BEAST 2: a software platform for Bayesian evolutionary analysis. *PLoS Comput Biol*. 10(4):e1003537.

Bull RA, Luciani F, McElroy K, Gaudieri S, Pham ST, Chopra A, Cameron B, Maher L, Dore GJ, White PA, et al. 2011. Sequential bottlenecks drive viral evolution in early acute hepatitis C virus infection. *PLoS Pathog*. 7(9):e1002243.

De Maio N, Wu CH, Wilson DJ. 2016. SCOTTI: efficient reconstruction of transmission within outbreaks with the structured coalescent. *PLoS Comput Biol*. 12(9):e1005130.

DeWitt WS 3rd, Mesin L, Victora GD, Minin VN, Matsen FA 4th. 2018. Using genotype abundance to improve phylogenetic inference. *Mol Biol Evol*. 35(5):1253–1265.

Didelot X, Fraser C, Gardy J, Colijn C. 2017. Genomic infectious disease epidemiology in partially sampled and ongoing outbreaks. *Mol Biol Evol*. 34(4):997–1007.

Didelot X, Gardy J, Colijn C. 2014. Bayesian inference of infectious disease transmission from whole-genome sequence data. *Mol Biol Evol*. 31(7):1869–1879.

Domingo E, Sheldon J, Perales C. 2012. Viral quasispecies evolution. *Microbiol Mol Biol Rev*. 76(2):159–216.

Drummond AJ, Rambaut A, Shapiro B, Pybus OG. 2005. Bayesian coalescent inference of past population dynamics from molecular sequences. *Mol Biol Evol*. 22(5):1185–1192.

Dudas G, Bedford T. 2019. The ability of single genes vs full genomes to resolve time and space in outbreak analysis. *BMC Evol Biol*. 19(1):17.

Eigen M, Schuster P. 1977. A principle of natural self-organization. *Naturwissenschaften* 64(11):541–565.
Faria NR, Azevedo RDSDS, Kraemer MUG, Souza R, Cunha MS, Hill SC, Theze J, Bonsall MB, Bowden TA, Rissanen I, et al. 2016. Zika virus in the Americas: early epidemiological and genetic findings. *Science* 352(6283):345–349.
Felsenstein J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol*. 17(6):368–376.
Ferguson NM, Cucunubá ZM, Dorigatti I, Nedjati-Gilani GL, Donnelly CA, Basáñez MG, Nouvellet P, Lessler J. 2016. Countering the Zika epidemic in Latin America. *Science* 353(6297):353–354.
Ferguson NM, Laydon D, Nedjati-Gilani G, et al. 2020. Impact of non-pharmaceutical interventions (NPIs) to reduce COVID-19 mortality and healthcare demand. Available from: https://spiral.imperial.ac.uk:8443/bitstream/10044/1/77482/14/2020-03-16-COVID19-Report-9.pdf (last accessed April 17, 2020).
González-Candelas F, López-Labrador FX, Bracho MA. 2011. Recombination in hepatitis C virus. *Viruses* 3(10):2006–2024.
Goodwin S, McPherson JD, McCombie WR. 2016. Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet*. 17(6):333–351.
Gray RR, Parker J, Lemey P, Salemi M, Katzourakis A, Pybus OG. 2011. The mode and tempo of hepatitis C virus evolution within and among hosts. *BMC Evol Biol*. 11(1):131.
Hall M, Woolhouse M, Rambaut A. 2015. Epidemic reconstruction in a phylogenetics framework: transmission trees as partitions of the node set. *PLoS Comput Biol*. 11(12):e1004613.
Hastings WK. 1970. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* 57(1):97–109.
Hué S, Pillay D, Clewley JP, Pybus OG. 2005. Genetic analysis reveals the complex structure of HIV-1 transmission within defined risk groups. *Proc Natl Acad Sci U S A*. 102(12):4425–4429.
Jukes TH, Cantor CR. 1969. Evolution of protein molecules. In: Munro HN, editor. Mammalian protein metabolism. Vol. 3. New York: Academic Press. p. 21–132.
Klinkenberg D, Backer JA, Didelot X, Colijn C, Wallinga J. 2017. Simultaneous inference of phylogenetic and transmission trees in infectious disease outbreaks. *PLoS Comput Biol*. 13(5):e1005495.
Malhotra R, Wu S, Mukhopadhyayand M, Rodrigo A, Poss M, Acharya R. 2016. Maximum Likelihood de novo reconstruction of viral populations using paired end sequencing data. *arXiv*:1502.04239.
McKinley TJ, Murcia PR, Gog JR, Varela M, Wood JL. 2011. A Bayesian approach to analyse genetic variation within RNA viral populations. *PLoS Comput Biol*. 7(3):e1002027.
Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E. 1953. Equation of state calculations by fast computing machines. *J Chem Phys*. 21(6):1087–1092.
Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, Lanfear R. 2020. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol Biol Evol*. 37(5):1530–1534.
Montoya V, Olmstead A, Tang P, Cook D, Janjua N, Grebely J, Jacka B, Poon AF, Krajden M. 2016. Deep sequencing increases hepatitis C virus phylogenetic cluster detection compared to Sanger sequencing. *Infect Genet Evol*. 43:329–337.
Pandit A, de Boer RJ. 2014. Reliable reconstruction of HIV-1 whole genome haplotypes reveals clonal interference and genetic hitchhiking among immune escape variants. *Retrovirology* 11(1):56.
Perelson AS, Neumann AU, Markowitz M, Leonard JM, Ho DD. 1996. HIV-1 dynamics in vivo: virion clearance rate, infected cell life-span, and viral generation time. *Science* 271(5255):1582–1586.
Poon AF, McGovern RA, Mo T, Knapp DJ, Brenner B, Routy JP, Wainberg MA, Harrigan PR. 2011. Dates of HIV infection can be estimated for seroprevalent patients by coalescent analysis of serial next-generation sequencing data. *AIDS* 25(16):2019–2026.
Poon AFY, Swenson LC, Bunnik EM, Edo-Matas D, Schuitemaker H, van't Wout AB, Harrigan PR. 2012. Reconstructing the dynamics of HIV evolution within hosts from serial deep sequence data. *PLoS Comput Biol*. 8(11):e1002753.
Prosperi MC, Yin L, Nolan DJ, Lowe AD, Goodenow MM, Salemi M. 2013. Empirical validation of viral quasispecies assembly algorithms: state-of-the-art and challenges. *Sci Rep*. 3(1):2837.
Pybus OG, Rambaut A. 2009. Evolutionary analysis of the dynamics of viral infectious disease. *Nat Rev Genet*. 10(8):540–550.
Recarey R, Cristina J. 2014. Bayesian coalescent analysis of the intra-host evolution of hepatitis C virus: memory genomes and clinical implications. *Nat Sci*. 6(9):615–627.
Ribeiro RM, Li H, Wang S, Stoddard MB, Learn GH, Korber BT, Bhattacharya T, Guedj J, Parrish EH, Hahn BH, et al. 2012. Quantifying the diversification of hepatitis C virus (HCV) during primary infection: estimates of the in vivo mutation rate. *PLoS Pathog*. 8(8):e1002881. (11 co-authors).
Romero-Severson EO, Bulla I, Leitner T. 2016. Phylogenetically resolving epidemiologic linkage. *Proc Natl Acad Sci U S A*. 113(10):2690–2695.
Schirmer M, Sloan WT, Quince C. 2014. Benchmarking of viral haplotype reconstruction programmes: an overview of the capacities and limitations of currently available programmes. *Brief Bioinform*. 15(3):431–442.
Schuster SC. 2008. Next-generation sequencing transforms today's biology. *Nat Methods*. 5(1):16–18.
Simon-Loriere E, Holmes EC. 2011. Why do RNA viruses recombine? *Nat Rev Microbiol*. 9(8):617–626.
Smyth RP, Schlub TE, Grimm AJ, Waugh C, Ellenberg P, Chopra A, Mallal S, Cromer D, Mak J, Davenport MP, et al. 2014. Identifying recombination hot spots in the HIV-1 genome. *J Virol*. 88(5):2891–2902.
Stadler T, Kühnert D, Bonhoeffer S, Drummond AJ. 2013. Birth–death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV). *Proc Natl Acad Sci U S A*. 110(1):228–233.
Stadler T, Kühnert D, Rasmussen DA, du Plessis L. 2014. Insights into the early epidemic spread of Ebola in Sierra Leone provided by viral sequence data. *PLoS Curr*. doi: 10.1371/currents.outbreaks.02bc6d927ecee7bbd33532ec8ba6a25f.
Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30(9):1312–1313.
Suchard MA, Rambaut A. 2009. Many-core algorithms for statistical phylogenetics. *Bioinformatics* 25(11):1370–1376.
Töpfer A, Marschall T, Bull RA, Luciani F, Schönhuth A, Beerenwinkel N. 2014. Viral quasispecies assembly via maximal clique enumeration. *PLoS Comput Biol*. 10(3):e1003515.
Töpfer A, Zagordi O, Prabhakaran S, Roth V, Halperin E, Beerenwinkel N. 2013. Probabilistic inference of viral quasispecies subject to recombination. *J Comput Biol*. 20(2):113–123.
Vaughan TG, Welch D, Drummond AJ, Biggs PJ, George T, French NP. 2017. Inferring ancestral recombination graphs from bacterial genomic data. *Genetics* 205(2):857–870.
Volz E, Pond S. 2014. Phylodynamic analysis of Ebola virus in the 2014 Sierra Leone epidemic. *PLoS Curr*. doi: 10.1371/currents.outbreaks.6f7025f1271821d4c815385b08f5f80e.
Volz EM, Romero-Severson E, Leitner T. 2017. Phylodynamic inference across epidemic scales. *Mol Biol Evol*. 34(5):1276–1288.
Vrancken B, Rambaut A, Suchard MA, Drummond A, Baele G, Derdelinckx I, Van Wijngaerden E, Vandamme AM, Van Laethem K, Lemey P. 2014. The genealogical population dynamics of HIV-1 in a large transmission chain: bridging within and among host evolutionary rates. *PLoS Comput Biol*. 10(4):e1003505.
Vrancken B, Trovão NS, Baele G, Van Wijngaerden E, Vandamme AM, Van Laethem K, Lemey P. 2016. Quantifying next generation sequencing sample pre-processing bias in HIV-1 complete genome sequencing. *Viruses* 8(1):12.
Wei X, Ghosh SK, Taylor ME, Johnson VA, Emini EA, Deutsch P, Lifson JD, Bonhoeffer S, Nowak MA, Hahn BH, et al. 1995. Viral dynamics in human immunodeficiency virus type 1 infection. *Nature* 373(6510):117–122.
Wilke CO. 2005. Quasispecies theory in the context of population genetics. *BMC Evol Biol*. 5(1):44.
Worby CJ, Lipsitch M, Hanage WP. 2014. Within-host bacterial diversity hinders accurate reconstruction of transmission networks from genomic distance data. *PLoS Comput Biol*. 10(3):e1003549.

Wu NC, De La Cruz J, Al-Mawsawi LQ, Olson CA, Qi H, Luan HH, Nguyen N, Du Y, Le S, Wu T-T, et al. 2014. HIV-1 quasispecies delineation by tag linkage deep sequencing. *PLoS One* 9(5):e97505.

Ypma RJ, van Ballegooijen WM, Wallinga J. 2013. Relating phylogenetic trees to transmission trees of infectious disease outbreaks. *Genetics* 195(3):1055–1062.

Zagordi O, Bhattacharya A, Eriksson N, Beerenwinkel N. 2011. ShoRAH: estimating the genetic diversity of a mixed sample from next-generation sequencing data. *BMC Bioinformatics* 12(1):119.

Zanini F, Brodin J, Thebo L, Lanz C, Bratt G, Albert J, Neher RA. 2015. Population genomics of intrapatient HIV-1 evolution. *Elife* 4:e11282.