Fast and accurate maximum-likelihood estimation of Birth-Death Exposed-Infectious epidemiological model from phylogenetic trees

Anna Zhukova[a,b,c,d,1], **Frédéric Hecht**[e], **Yvon Maday**[e,f], **and Olivier Gascuel**[a,g,1]

[a]Unité Bioinformatique Evolutive, Département de Biologie Computationelle, Institut Pasteur, Université de Paris, 75015 Paris, France; [b]Hub Bioinformatique et Biostatistique, Département de Biologie Computationelle, Institut Pasteur, Université de Paris, 75015 Paris, France; [c]Epidémiologie & modélisation de la résistance aux antimicrobiens, Institut Pasteur, Université de Paris, 75015 Paris, France; [d]Université Paris-Saclay, UVSQ, Inserm, CESP, 94807, Villejuif, France; [e]Sorbonne Université, CNRS, Université Paris Cité, Laboratoire Jacques-Louis Lions (LJLL), F-75005 Paris, France; [f]Institut Universitaire de France; [g]Institut de Systématique, Evolution, Biodiversité (ISYEB) - URM 7205 CNRS, Muséum National d'Histoire Naturelle, SU, EPHE & UA, 75005 Paris, France

This manuscript was compiled on August 1, 2022

The birth-death exposed-infectious (BDEI) model describes the transmission of pathogens featuring an incubation period (when the host is already infected but not yet infectious), for example Ebola and SARS-CoV-2. In a phylodynamics framework, it serves to infer such epidemiological parameters as the basic reproduction number R_0, the incubation period and the infectious time from a phylogenetic tree (a genealogy of pathogen sequences). With constantly growing sequencing data, the BDEI model should be extremely useful for unravelling information on pathogen epidemics.

However, existing implementations of this model in a phylodynamic framework have not yet caught up with the sequencing speed. While the accuracy of estimations should increase with data set size, existing BDEI implementations are limited to medium data sets of up to 500 samples, for both computing time and numerical instability reasons.

We improve accuracy and drastically reduce computing time for the BDEI model by rewriting its differential equations in a highly parallelizable way, and by using a combination of numerical analysis methods for their efficient resolution. Our implementation takes one minute on a phylogenetic tree of 10 000 samples. We compare our parameter estimator to the existing implementations on simulated data. Results show that we are not only much faster (50 000 times), but also more accurate. An application of our method to the 2014 Ebola epidemic in Sierra-Leone is also convincing, with very fast calculation and precise estimates. Our BDEI estimator should become an important tool for routine epidemiological surveillance. It is available at github.com/evolbioinfo/BDEI.

Phylodynamics | Epidemiology | Mathematical modelling | Ordinary Differential Equations | Ebola

The interaction of epidemiological and evolutionary processes leaves a footprint in pathogen genomes. Phylodynamics leverages this footprint to estimate epidemiological parameters (1, 2). It relies on models that bridge the gap between traditional epidemiology and sequence data by estimating parameters like the basic reproduction number, R_0, from topology and branch lengths of pathogen phylogenies (i.e. genealogies of the pathogen population, approximating the transmission trees) combined with metadata on the samples. This is particularly useful for emerging epidemics, for which not enough data (e.g. incidence curves) might yet be gathered for accurate estimations with classical epidemiological methods: Rapidly growing genetic data coupled with phylodynamic estimations can provide valuable insights at an early stage of the epidemic spread and help prevent it (e.g. accurate estimation of the infectious period is crucial for adjusting policies for self-isolation).

Phylodynamic models can be classified into two main families: coalescent (3–5) and birth-death (BD) (6–9). Coalescent models are often preferred for estimating deterministic population dynamics, however BD models are better adapted for highly stochastic processes, such as the

Significance Statement

Phylodynamics uses pathogen genomes as a source of information for epidemiological parameter inference. With constantly growing genome sequence availability, phylodynamics has a high potential for shedding light on epidemics, especially in the beginning, when classical epidemiological data (e.g. incidence curves) are yet limited. However due to the high complexity of differential equations used in phylodynamic models, current implementations suffer from numerical instability and are only applicable to datasets of limited size (<500 sequences).

We solve this computational bottleneck for the birth-death exposed-infectious model, which describes the transmission of pathogens with an incubation period (e.g. Ebola, SARS-CoV-2). Our fast and accurate estimator is applicable to very large datasets (10 000 samples) permitting phylodynamics to catch up with constantly growing pathogen sequencing efforts.

OG initiated the project; AZ conceived the new master equation representation, extension to forests, applications to Ebola and simulated data, and coordinated the project; FH and YM conceived the numerical approach for fast and efficient parameter optimisation; FH implemented the numerical approach; AZ wrote the manuscript, OG, FH and YM edited the manuscript; all authors discussed the intermediate and final results and the manuscript.

The authors declare no competing interests.

[1]To whom correspondence should be addressed. E-mail: anna.zhukova@pasteur.fr (AZ), olivier.gascuel@mnhn.fr (OG)

dynamics of emerging pathogens (10). In BD models, births represent pathogen transmission events, while deaths correspond to becoming non-infectious (e.g. due to healing, self-isolation, starting a treatment, or death). Models of the BD family are phylodynamic analogies of compartmental models in classical epidemiology (e.g. SIR, Susceptible-Infectious-Removed). Many extensions of the classical BD model with incomplete sampling (BDS (8)) were developed over time, including multi-type birth-death (MTBD) models (11), which add a population structure to the classical birth-death process by allowing for different types of individuals. A particularly useful representative of the MTBD family is the birth-death exposed-infectious (BDEI) model (12), which was designed for pathogens featuring an incubation period between the moments of infection and of becoming infectious, e.g. Ebola and SARS-CoV-2. It is closely related to the Susceptible-Exposed-Infectious-Recovered (SEIR) model (13), widely used in classical epidemiology.

In MTBD framework, the evolution of a transmission tree is described with a system of master differential equations with respect to global time. The model parameters can be estimated with maximum-likelihood (11) or Bayesian methods (15) by exploring the likelihood (or posterior probability) landscape of trees. However, the closed form solution of the master equations exists only for the initial BDS model, while for its extensions (like MTBD) the master equations for likelihood calculation need to be resolved with numerical methods. The complexity of the master equations and their initial conditions (which recursively depend on the tree evolution later in time), make their numerical resolution challenging and time consuming (16, 17).

The trade-off between the complexity of the biological questions a model can address, its computational speed and the size of the input data set is crucial in phylodynamics. On one hand, a denser sampling should improve the accuracy of parameter estimations with complex models, on the other hand it leads to larger data sets (thousands of samples), while computational issues often limit model applicability to medium or small ones (hundreds of samples). Calculations become time-consuming and numerically challenging (e.g. due to underflow issues) as tree size increases, resulting in numerical instability and inaccuracy (16, 17). Existing likelihood-based implementations of the BDEI model (11, 15, 16) can handle trees of medium size (hundreds of samples). In (17) we proposed PhyloDeep, a likelihood-free deep-learning-based solution to the numerical instability issue. While being very efficient at the prediction stage, this approach however requires a computationally heavy training stage: Millions of trees covering a wide parameter range (where the real data is expected to fall) need to be simulated and used for training the deep learning predictor.

In this study we fix the computational bottleneck and extend the applicability of the BDEI model by proposing a likelihood-based approach that improves the accuracy and reduces the likelihood computation time. For that we (i) identify a subclass of MTBD models (including the BDEI model) for which the likelihood formulae can be expressed in a highly parallelizable way, which avoids underflow issues; and (ii) develop targeted numerical analysis methods permitting accurate and fast resolution of the equations involved in the computation of the likelihood for the BDEI model. We show the accuracy and speed of our parameter estimator PyBDEI on simulated data and compare it to the gold standard Bayesian tool BEAST2 (15) and the deep-learning-based tool PhyloDeep (17). We find that our approach outperforms the competitors and makes the BDEI model applicable to very large data sets. Lastly, we apply PyBDEI to infer the epidemiological parameters that shaped the Ebola epidemic in Sierra-Leone in 2014. Our estimator is freely available from github.com/evolbioinfo/bdei.

1. Results

The BDEI model. In a pathogen transmission tree $\mathscr{T}$ (approximated by a time-scaled pathogen phylogeny) the tips represent sampled pathogens, patient state transitions occur along the branches, and bifurcations (i.e. internal nodes) correspond to pathogen transmissions (Fig. 1). The tree branch lengths are measured in units of time, where T is the time that passed between the tree root (the beginning of the (sub-)epidemic) and the last sampled tip.

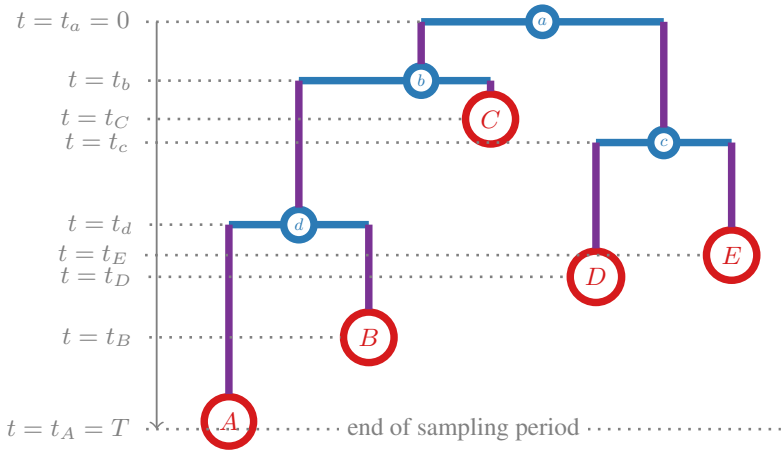

Fig. 1. A transmission tree $\mathscr{T}$ with $n = 5$ external nodes (i.e. tips, which correspond to sampling events: A, B, C, D, E), $n - 1 = 4$ internal nodes (which correspond to transmissions: a, b, c, d) and $2n - 2 = 8$ branches (plus the root branch of zero length). Time t starts at the root of the tree ($t = 0$) and goes till the last sampled tip. The times of the nodes are shown on the left, e.g. t_B is the time of tip B (when B's pathogen was sampled). T corresponds to the end of the sampling period (when the most recent tip, A, was sampled).

The BDEI model (Fig. 2) has two possible states:

- infectious I, an individual who can transmit the pathogen further or get removed from the system (with potential sampling);
- exposed E, an individual who is already infected but not yet infectious (cannot transmit), and will eventually become infectious.

At the moment of a transmission, the transmitter is always in state I, while the recipient is in state E. However, we typically do not have the information to distinguish a transmitter from a recipient in a phylogenetic tree (which approximates the transmission tree), and hence

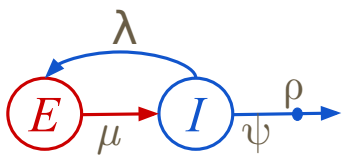

Fig. 2. The BDEI model. An individual in exposed state E becomes infectious at a rate μ. An infectious individual I transmits the pathogen at a rate λ (hence creating a new exposed individual E), and gets removed at a rate ψ (decreasing the number of infectious individuals I). Upon removal, the individual's pathogen might be observed with a probability ρ. Note, that the BDEI model does not include a susceptible state S (as for example SEIR) and makes the assumption that the susceptible population is unlimited (as for example in the beginning of an epidemic, or when the removed individuals could get reinfected).

consider both possibilities. While inner branches finish with a transmission event, tip branches finish with a sampling: After sampling, the individual (and their pathogen) exits the study. In the BDEI model, we assume that only the individuals in state I can be detected and sampled: For instance, for many pathogens with an incubation period, the detection is triggered by the onset of symptoms, which in turn happens in the infectious state. Hence all the tree tips are in state I. However, the sampling is incomplete: an infectious individual may be removed from the system without being sampled (i.e. unobserved in the transmission tree), for example due to healing.

The BDEI model permits the inference of such important epidemiological parameters as the basic reproduction number R_0 (the expected number of individuals directly infected by an infectious case), the incubation period (time between the infection and becoming infectious), and the infectious time (time during which the individual can further spread the epidemic). They can be expressed via 3 exponential rates:

- μ – becoming infectious rate, corresponding to a state transition from E to I;
- λ – transmission rate, from a transmitter in state I to a newly infected recipient, whose state is E;
- ψ – removal rate, corresponding to becoming non-infectious of an individual in state I (e.g. due to healing, death or starting a treatment).

The epidemiological parameters can be estimated from the rate parameters as: $R_0 = \frac{\lambda}{\psi}$, the incubation period is $\frac{1}{\mu}$, and the infectious time is $\frac{1}{\psi}$. The fourth model parameter is the sampling probability ρ – the probability to sample the pathogen (and therefore observe it as a tip of the tree) upon removal of an individual in state I. This parameter is needed as not all the removed infectious individuals are sampled, as, for example, asymptomatic persons who healed.

The BDEI model, as an extension of the BDS model, is asymptotically unidentifiable (see Remark 3.4 in (8)), but to become identifiable it requires one of its parameters to be fixed. In practice, it is often the sampling probability ρ, as it may be approximated from epidemiological data (e.g. the proportion of sampled cases among the declared ones) or the infectious time $\frac{1}{\psi}$ (estimated from observations of infected cases).

Master equations. In the MTBD framework, time goes backward from the last sampling event (the most recent tip in the tree) till the beginning of the epidemic. Stadler *et al.* (11) developed master equations for MTBD models. In System [1] we show the special case of these equations that corresponds to the BDEI model, however presenting them with the time t going forward from the time of the root ($t = 0$) to the time of the last sampled tip ($t = T$). These equations describe the probability density functions (PDFs) $P_s^{(i)}(t)$ of an individual evolving as observed in the tree, starting at time t in state $s \in \{E, I\}$ on a branch connecting a node i to its parent, and evolving till the end of the sampling period. The initial conditions are defined at time $t = t_i$ (i.e. at the node i). Note that the node i can correspond either to a transmission (an internal node) or a sampling (a tip), however in both cases it is in state I as only infectious individuals can transmit or get sampled. To account for incomplete sampling, the system also includes the probabilities $U_s(t)$ of evolving unobserved till the time T, starting at time t in state s.

$$
\begin{cases}
\dot{P}_E^{(i)}(t) = & \mu P_E^{(i)}(t) \leftarrow \textit{no event in the next infinitesimal time } \Delta t \\
& -\mu P_I^{(i)}(t) \leftarrow \textit{becoming infectious, followed by evolution from state } I \\
\dot{P}_I^{(i)}(t) = & (\lambda + \psi) P_I^{(i)}(t) \leftarrow \textit{no event in the next infinitesimal time } \Delta t \\
& -\lambda P_I^{(i)}(t) U_E(t) \leftarrow \textit{transmission, where the recipient subtree stayed unsampled} \\
& -\lambda P_E^{(i)}(t) U_I(t) \leftarrow \textit{transmission, where the donor subtree stayed unsampled} \\
P_I^{(i)}(t_i) = & C^{(i)} = \begin{cases} \psi\rho, & \textit{if } i \textit{ is a sampled tip} \\ \lambda\left(P_I^{(\textit{left child of } i)}(t_i) P_E^{(\textit{right child of } i)}(t_i) + P_E^{(\textit{left child of } i)}(t_i) P_I^{(\textit{right child of } i)}(t_i)\right), & \textit{if } i \textit{ is an internal node (transmission)} \end{cases} \\
P_E^{(i)}(t_i) = & 0 \leftarrow \textit{an individual in state } E \textit{ cannot become infectious } (I) \textit{ over time 0} \\
\dot{U}_E(t) = & \mu U_E(t) \leftarrow \textit{no event in the next infinitesimal time } \Delta t \\
& -\mu U_I(t) \leftarrow \textit{becoming infectious, followed by unsampled evolution from state } I \\
\dot{U}_I(t) = & (\lambda + \psi) U_I(t) \leftarrow \textit{no event in the next infinitesimal time } \Delta t \\
& -\lambda U_I(t) U_E(t) \leftarrow \textit{transmission, followed by unsampled evolutions of both the donor and the recipient subtrees} \\
& -\psi(1 - \rho) \leftarrow \textit{removal without sampling} \\
U_E(T) = & 1 \leftarrow \textit{the probability to stay unsampled over time 0 is 1} \\
U_I(T) = & 1 \leftarrow \textit{the probability to stay unsampled over time 0 is 1}
\end{cases}
\qquad [1]
$$

Tree likelihood. The likelihood of a tree $\mathscr{T}$ for given parameter values $\Theta = \{\mu, \lambda, \psi, \rho\}$ is then calculated as the PDF at time $t = 0$ (root):

$$L(\mathscr{T}|\Theta) = P_I^{(root)}(0) \qquad [2]$$

It *recursively* depends on the PDFs of the child node branches via the initial condition $C^{(root)}$, and hence is calculated with a pruning algorithm (18) while climbing the tree from tips till the root. Therefore when parallelized to maximum, it still requires $O(h_{\mathscr{T}})$ consecutive steps, where $h_{\mathscr{T}}$ stands for the height of the tree $\mathscr{T}$ and depends on its topology: (balanced tree) $log(n) \leq h_{\mathscr{T}} \leq n$ (ladder-like tree). At each step System [1] needs to be resolved for the corresponding nodes. Moreover, the values of $P_s^{(i)}(t)$ at internal nodes and their initial conditions $C^{(i)}$ progressively become smaller as getting deeper in the tree, due to successive additions and multiplications of the PDF values. In trees with many tips, this might lead to numerical underflow, and hence such measures as rescaling need to be taken (16, 19, 20).

Avoiding numerical problems and parallelizing calculations. In this subsection we introduce a way to rewrite PDFs in System [1] that permits (1) obtaining simpler initial conditions to avoid potential numerical issues during resolution of equations; and (2) removing recursion and resolving equations for each tree node in parallel, hence speeding up the calculations.

System [1] (as well as System [12] corresponding to a general MTBD model, see Materials and Methods) has several properties. First of all, its subsystem that defines unobserved probabilities ($U_E(t)$ and $U_I(t)$ for the BDEI model) is self-defined, and hence can be calculated independently from the rest. Secondly, in the subsystem that defines observed PDFs ($P_I^{(i)}(t)$ and $P_E^{(i)}(t)$) the right-hand side of the differential equations is a sum whose elements are linear with respect to either $P_I^{(i)}(t)$ or $P_E^{(i)}(t)$, and this sum does not contain a free term. This condition implies that if we rescale $P_I^{(i)}(t)$ and $P_E^{(i)}(t)$ by a common factor, the differential equations will not change. Moreover, as the initial condition for $P_E^{(i)}(t)$ (at $t = t_i$) is zero, the rescaling will only change the initial condition for $P_I^{(i)}(t)$ (at $t = t_i$).

Let us define $p_s^{(i)}(t)$, where $s \in \{E, I\}$ as:

$$\begin{cases} p_I^{(i)}(t) &= P_I^{(i)}(t)/C^{(i)} \\ p_E^{(i)}(t) &= P_E^{(i)}(t)/C^{(i)} \end{cases}. \qquad [3]$$

Then the differential equations for $p_s^{(i)}(t)$ will only differ from those for $P_s^{(i)}(t)$ in the initial condition for $s = I$, which is $p_I^{(i)}(t_i) = 1$. Conceptually, $p_s^{(i)}(t)$ is a PDF of an individual evolving as on an observed branch that connects a node i to its parent, starting at time t in state s on this branch and finishing at time t_i in state I (without taking into account i's subtree and the event at node i).

Solving the master equations for $p_s^{(i)}(t)$ instead of $P_s^{(i)}(t)$ permits us both (1) to avoid numerical issues that could arise from very small values of the initial condition of $P_I^{(i)}(t)$, which is particularly pertinent for large trees; and (2) to remove the recursive dependency between child and parent nodes, thus permitting their parallel calculation. The calculation of $p_s^{(i)}(t)$ can be done in parallel for each node i (i.e. constant number of master equation resolutions when parallelized on $2n - 2$ (number of non-root tree nodes) cores). This PDF reconditioning technique can be generalized to any MTBD model, as we explain in Materials and Methods.

Finally, as all the internal node states of the tree are known (I), the state of a parent node does not depend on its child states, and hence we can write the tree likelihood in a non-recursive logarithmic form:

$$\begin{aligned} logL(\mathscr{T}|\Theta) = {} & nlog(\psi\rho) && \leftarrow \textit{sampling of } n \textit{ tips} \\ & + (n-1)log\lambda && \leftarrow n-1 \textit{ transmission events} \\ & + \sum_{i \in \textit{internal nodes}} log\Big(p_I^{(\textit{left child of } i)}(t_i)p_E^{(\textit{right child of } i)}(t_i) && \leftarrow \textit{child branch evolutions} \\ & \qquad + p_E^{(\textit{left child of } i)}(t_i)p_I^{(\textit{right child of } i)}(t_i)\Big) && \textit{for each internal node} \end{aligned} \qquad [4]$$

In Materials and Methods we show the equivalence between Eq. [2] and Eq. [4], and explain that such a non-recursive log-likelihood equation can be obtained for any MTBD model on a tree with known internal node states (Eq. [13]). Logarithmic representation helps avoid underflow issues (we sum up log values instead of multiplying very small PDF values).

Overall, the PDF reconditioning technique can be applied to any model of the MTBD family, and facilitates its parameter estimation by separating ODE resolution (non-recursive and parallelizable) from likelihood calculation (recursive, but negligible in time cost compared to ODE resolution, see Eq. [15] in Materials and Methods). Recursive likelihood calculation can be performed with a standard pruning algorithm and rescaling techniques to control for potential underflow. For parameter estimation on trees with known node states (e.g. from metadata, or because they were generated by an MTBD process in which only one state can transmit or get sampled, like the BDEI model), tree likelihood can be calculated with a non-recursive formula in a logarithmic form (Eq. [4] and Eq. [13]), avoiding underflow.

Forests. In some cases, the assumption that a (sub-)epidemic started with one infected individual might be too constraining. For instance, there could be multiple pathogen introductions to a country of interest (e.g. while in China the SARS-CoV-2 epidemic is commonly assumed to have started with one case, there were multiple independent introductions to other countries (21)). Another example is a change of health policies leading to a change in parameter values (e.g. sampling). Such a change corresponds to a new stage of the epidemic, starting from several infected cases from the previous stage. In Bayesian settings, the situations when the system behaviour (and parameters) change over time, are modelled via skyline methods. Stadler *et al.* (22) developed the one-state Bayesian birth-death skyline plot that divides the time into intervals and allows for different piecewise constant rates on them. Kühnert *et al.* (23) combined the MTBD model with the skyline to allow for both piecewise-constant rate changes over time and multiple individual types. The skyline approach therefore relies on a single tree, but estimates a separate set of parameters for each time interval, all under the same model. As the number of parameters increases with multiple skyline intervals, the BDEI-skyline model therefore requires more data for their accurate estimation, more computational time and is more prone to numerical instability than the classical BDEI model.

We propose a simpler alternative, where the (sub-)epidemic starts with multiple individuals (not necessarily at the same time) and leads to f observed trees with n sampled tips in total. If all the trees started at the same time (e.g. due to health policy change), the information on

the number of declared cases (m) at the start of the (sub-)epidemic might be available. If this number is larger than the number of observed trees ($m > f$), it implies that the other $u = m - f$ trees stay unobserved as none of their tips got sampled, which can be incorporated in the likelihood calculation. Forest likelihood formula (see Eq. [6] in Materials and Methods) hence combines the likelihoods of f observed and u hidden trees. Tree likelihood formula [4] is its special case, where $f = 1$ and $u = 0$.

Using forests allows to estimate the BDEI model parameters on the last skyline interval without the restriction that the epidemic followed the same model before this interval (i.e. the top part of the tree, which includes the common ancestors of the forest roots). It reduces the number of parameters to those of the last interval, and informs the inference with the number of declared cases at the start of the sub-epidemic. It also permits estimation of parameters for a (sub-)epidemic that started with several individuals but not at the same time (e.g. multiple introductions to a country).

Efficient parameter and CI estimation. We estimate the BDEI model parameters $\Theta = (\mu, \lambda, \psi, \rho) \in \mathbb{R}^4$ for a forest $\mathscr{F}$ (comprising $f \geq 1$ observed and $u \geq 0$ unobserved trees) in the maximum-likelihood framework, where one of the parameters in Θ is fixed. The estimation starts with a preprocessing step of reading the input trees, calculating the time t_i at each node i and memorising the association between each node and its child nodes: $i \rightarrow (\textit{left child of } i, \textit{right child of } i)$. This step requires one tree traversal, and its time is negligible with respect to the numerical resolution of the differential equations that is performed in the next steps.

The estimation then proceeds with a search for the optimal parameter set $\Theta_{opt} = \underset{\Theta_k \in \mathcal{Q}}{\arg\max}\, L(\mathscr{F}|\Theta_k)$, where $\mathcal{Q}$ is the set of admissible parameter values. We use the globally-convergent method of moving asymptotes (24) for the optimisation. At each optimisation step k the corresponding likelihood $L(\mathscr{F}|\Theta_k)$ needs to be calculated, which implies calculating $p_I^{(\textit{left child of } i)}(t_i), p_E^{(\textit{left child of } i)}(t_i), p_I^{(\textit{right child of } i)}(t_i), p_E^{(\textit{right child of } i)}(t_i)$ for each of the $n - f$ observed internal nodes of $\mathscr{F}$, and combining them as in the forest likelihood formula (Eq. [6]). The reconditioned version of master equations [1] for the parameter values Θ_k can be resolved in parallel for each of the $2n - f$ observed forest nodes (differing in the times t_i in their initial conditions ${p_I}^{(i)}(t_i) = 1$).

To resolve these master equations numerically, we start by separating the self-defined subsystem for the unknowns $U_E(t)$ and $U_I(t)$ from the rest of System [1] to calculate it independently. Taking into account the fact that the equations in System [1] are either linear (for $p_E^{(i)}(t)$ and $p_I^{(i)}(t)$) or quadratic (for $U_E(t)$ and $U_I(t)$) we note that the use of an implicit scheme is simple. We chose implicit schemes with an automatic computation of the time step, such that the error is less than a given tolerance. In our implementation, we used the implicit Euler scheme (26) for solving the linear equations and the Crank-Nicolson implicit scheme (25) for the non-linear ones. This allows to avoid possible stability time restrictions and only choose the time steps for precision.

Once the optimal parameter values are found, we calculate their confidence intervals (CIs) using Wilks' method (27). For each of the non-fixed parameters $p \in \Theta$, we calculate its 95%-CI as including values $\tilde{p}$ such that $logL(\mathscr{F}|\Theta_{opt|p=\tilde{p}}) > logL(\mathscr{F}|\Theta_{opt}) - \chi^2_1(0.95)/2$, where $\Theta_{opt|p=\tilde{p}}$ corresponds to the maximum-likelihood value of the other two non-fixed parameters when $p = \tilde{p}$.

Performance on simulated data and comparison to other tools. To assess the performance of our maximum-likelihood estimator (which we called PyBDEI), we used the simulated data from (17), where we generated 10 000 medium trees with 200–500 tips under the BDEI model, with the parameter values sampled uniformly at random within the following boundaries: incubation period $\frac{1}{\mu} \in [0.2, 50]$, basic reproduction number $R_0 = \frac{\lambda}{\psi} \in [1, 5]$, infectious period $\frac{1}{\psi} \in [1, 10]$. Out of these 10 000 trees, 100 were randomly selected and evaluated with the gold standard Bayesian method BEAST2 (15) and the deep learning-based estimator PhyloDeep (detailed configurations are described in Materials and Methods). Additionally 100 large trees (5 000–10 000 tips) were generated for the same parameter values as the 100 selected medium trees, and assessed with PhyloDeep in (17). PhyloDeep's maximal pre-trained tree size is 500 tips, however for larger trees it estimates BDEI parameters by (1) extracting the largest non-intersecting set of subtrees of sizes covered by the pre-trained set (50–500 tips), (2) estimating parameters on each of the subtrees independently, and (3) averaging each parameter's estimate over the subtrees (weighted by subtree sizes).

We applied PyBDEI to these data sets (for the large data set, both to full trees and to forests of subtrees produced by PhyloDeep), and compared the results to those reported for BEAST2 and PhyloDeep in (17). We calculated the relative error (normalized distance between the estimated and the target values: $\frac{|estimated - target|}{target}$) and the relative bias ($\frac{estimated - target}{target}$) for each parameter on each tree. Average relative errors for PyBDEI were $\leq 13\%$ on the medium trees and $\leq 2\%$ on the large trees (hence decreasing with the data set size, as expected), and well centred around zero (i.e. unbiased), as shown in Fig. 3. The relative CI width ($\frac{target_{95\%} - target_{5\%}}{target}$) also decreased: from ~ 0.5 on the medium data set to ~ 0.1 on the large one. The target values of rates μ, λ and ψ were within the estimated CIs in correspondingly 92%, 89%, and 98% of cases on the medium data set, and in 96%, 90% and 94% of cases on the large one.

In terms of accuracy, on the medium data set PyBDEI was at least as accurate as PhyloDeep and more accurate than BEAST2 ($p < 0.05$ for all the parameters but R_0, where all the methods performed in a comparable way, see Fig. 3). On the large data set BEAST2 was inapplicable due to computation times (57 CPU hours were already required for each medium tree, on average), while PyBDEI was more accurate than PhyloDeep, both using full trees and forests ($p < 0.01$ for all the parameters, see Fig. 3).

In terms of time, on the medium data set PyBDEI needed on average 4 seconds per tree on 1 CPU, and converged in 461 iterations. These times cannot be directly compared to BEAST2 times, as BEAST2 performs a Markov Chain Monte Carlo (MCMC) parameter space exploration instead of looking for the optimum, hence requires many more steps. BEAST2 required on average 57 CPU hours for 10^6 MCMC steps. Comparing time per iteration (which is roughly time per likelihood calculation), our optimiser required ~ 0.01 CPU seconds, while BEAST2 took one order of magnitude longer: ~ 0.2 CPU seconds. PhyloDeep took 0.2 CPU seconds per tree, which is faster than our method's time but does not include the training time of deep learning predictors (hundreds of hours). Neither can this value be converted into time per iteration as it is a likelihood-free deep learning-based method. To our knowledge, the only other available maximum-likelihood estimator for BDEI is implemented in the TreePar package (11). However, as it suffers from underflow issues for BDEI already on trees of medium size, its developers suggest using BEAST2 instead (private communication). The average time of PyBDEI convergence on the large

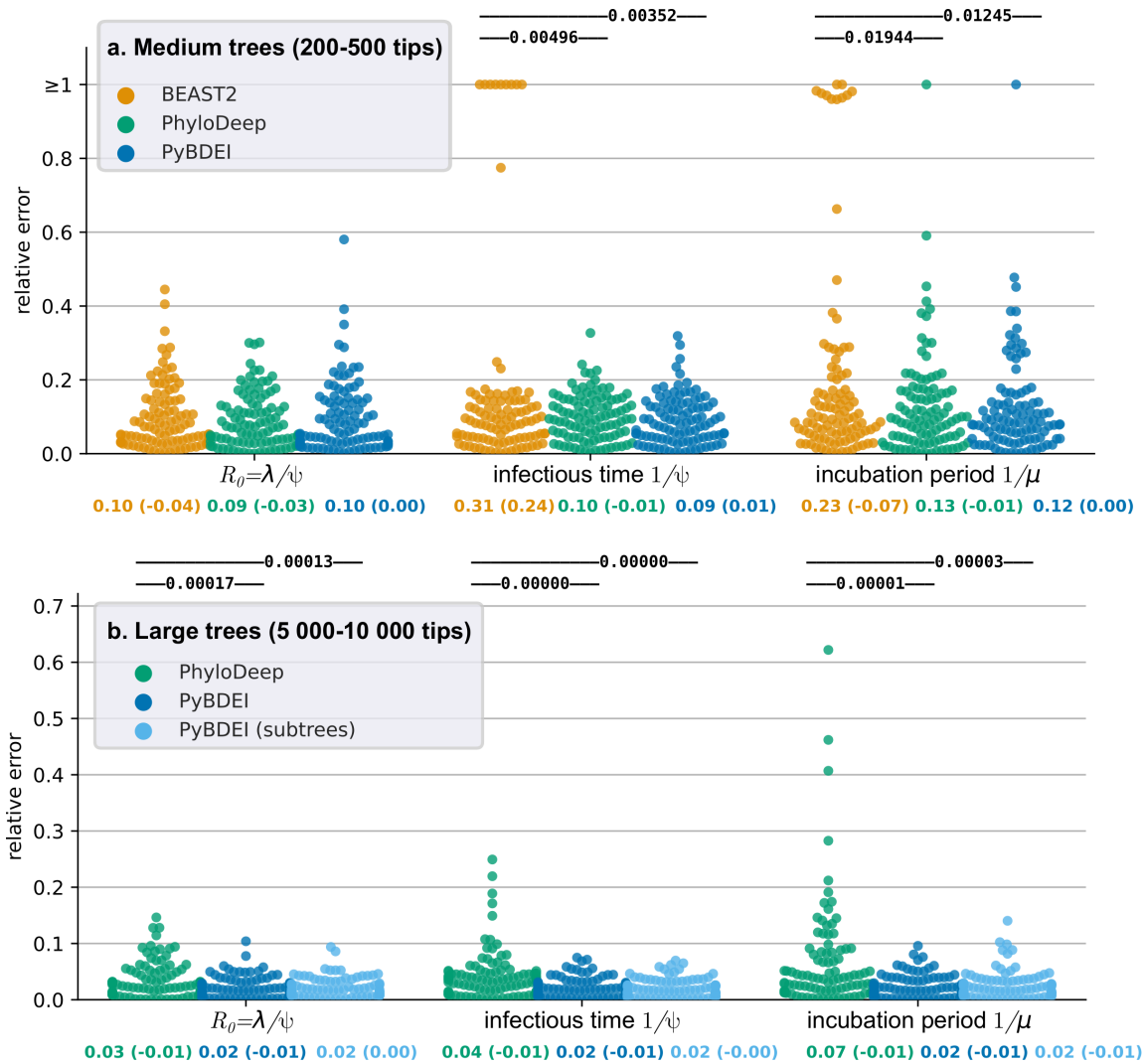

Fig. 3. Comparison of inference accuracy of different methods on the medium (200-500 tips, top) and large (5 000–10 000 tips, bottom) data sets. For the medium data set BEAST2 (in orange), PhyloDeep (in green) and our estimator (PyBDEI, in blue) are compared. For large trees ($>$ 500 tips), PhyloDeep extracts the largest non-intersecting set of subtrees of 50–500 tips, estimates parameters on each of the subtrees independently, and averages each parameter's estimate over the subtrees (weighted by subtree sizes). We assessed our method both on full trees (dark-blue) and forests of subtrees extracted by PhyloDeep (light-blue). We show the swarmplots (coloured by method) of relative errors for each test tree and parameter, which are measured as the normalized distance between the median a posteriori estimate by BEAST2 or a point estimate by PhyloDeep/PyBDEI and the real value. Average relative error (and in parentheses average bias, calculated on normalized values) are displayed for each parameter and method below their swarmplot. The accuracy of the methods is compared by a paired z-test; $p < 0.05$ are shown above each method pair; non-significant p-values are not shown.

data set was 111 seconds on 1 CPU, and required 492 iterations. Parallelization on 2 CPUs reduced it to 68 seconds (1.6 times faster). The speed up is close to the number of cores, which shows the efficiency of parallelization of ODE resolution despite the pre- and post-processing steps (tree reading, distribution of jobs between the threads, combining their results), which are always performed on one CPU. This suggests that our estimator will be easily applicable to much larger trees.

As the BDEI model requires one of the parameters to be fixed in order to become asymptomatically identifiable (8), we fixed ρ to the real value, both in (17) and in the comparison described above. However, to assess PyBDEI performance with other parameters fixed, we estimated parameters for trees in the large data set under three additional settings: with (1) μ, (2) λ, or (3) ψ fixed to its real value. The results are shown in Fig. S1. Average relative errors were $\leq 4\%$ for all parameters when μ was fixed to the real value, $\leq 3\%$ when λ was fixed, $\leq 2\%$ when ψ or ρ were fixed. For estimates of ρ we calculated absolute errors $|estimated - target|$ instead of relative ones $\frac{|estimated-target|}{target}$: their average was 0.02 for fixed μ or λ , and 0.01 for fixed ψ. Hence, the estimations can be successfully performed with any of the parameters being fixed, but fixing ψ might be particularly useful.

The assessment of our maximum-likelihood estimator on simulated data shows that it opens new possibilities to fast and efficient analyses of extremely large data sets, while being flexible with respect to parameter settings (e.g. the parameter to be fixed).

Application: Ebola in Sierra-Leone. Using PyBDEI, we analysed the 2014 Ebola epidemic in Sierra-Leone (SLE). Ebola virus features an incubation period (reported by WHO to take between 2 and 21 days, see who.int). Using statistical methods based on time series of reported Ebola cases, the incubation period of Ebola during the 2014 SLE epidemic was previously estimated to be around 10–11 days, the infectious

time around 4–5 days, and the reproduction number to decrease from around 2 in the beginning of the epidemic to values close to 1 by late 2014 (due to control measures) (28–30) (see also (31) for a review).

Sequence data could improve and complement these estimates. However, the existing phylodynamic study of these parameters was limited by the data set size: Stadler *et al.*(11) applied the BDEI model to the early spread of Ebola in SLE by analysing 72 Ebola samples from late May to mid June 2014 (sequences from Gire *et al.* (32)). They estimated the expected length of the infectious period to be 2.6 days (median; 95% HPD 1.2-7), and the incubation period of 4.9 days (median; 95% HPD 2.1-23.2).

To show the power of phylodynamic analyses on larger data sets, we took the 1 610-sequence alignment and metadata (sampling times and countries) that were used in the study by Dudas *et al.* (33), who analysed the factors that spread the 2014–2016 Ebola epidemic in West Africa. Using these data, we reconstructed a time-tree of the Ebola epidemic in West Africa, which we then used to extract a forest of time-subtrees representing the Ebola epidemic in SLE between July 30, 2014 (when the SLE government began to deploy troops to enforce quarantines according to news24.com) and September 12, 2015 (the last SLE sample in the data set). This was done to obtain a forest of subtrees with a homogenous health policy (after July 30). The details on the forest reconstruction are given in Materials and Methods. To check for robustness of the estimates, forest reconstruction was performed 10 times, obtaining slightly different trees.

We estimated the BDEI parameters on these 10 forests, fixing the total number of trees to $N = 533$ (the declared number of cases in SLE by July 31, 2014 from cdc.gov), and hence the number of unobserved trees to $N - k$, where k was the number of trees in the corresponding forest (k varied between 55 and 70). To check the robustness of the predictions with respect to N, we additionally estimated the parameter values assuming 50% more cases, i.e. $N = 800$.

As the BDEI model requires one of the parameters to be fixed for identifiability (8), we performed the estimations fixing the infectious period (to 2.6 and to 5 days, i.e. the estimates from the previous studies). The results are listed in Table S1: the effective reproduction number R_e was ~ 0.9 $[0.94 - 0.98]$ in all the settings and forests, suggesting a contained epidemic (which is in a good agreement with the quarantine measures and the end of the epidemic in early 2016), the incubation period was estimated between 9.7 and 14.4. These estimates are fully compatible with the previous studies and allow to narrow the previous incubation period estimate (2–23 days, non-specific to SLE epidemic) down to 10–14 days. The sampling probability ρ was estimated ~ 0.1 $[0.08 - 0.16]$. This number corresponds quite well to the proportion of cases represented by our forests (755–779) with respect to the total number of people infected during the SLE 2014 Ebola outbreak (8 704, as reported by WHO at afro.who.int): $760/8\,704 \approx 0.09$.

Overall, the analysis took ~ 1 hour for the forest reconstruction and < 5 seconds for the BDEI parameter estimation.

This application shows the advantages of PyBDEI not only in terms of calculation times, but also in terms of flexibility of input settings (parameter to be fixed, extracting information from multiple trees and using data on declared but unobserved cases).

2. Discussion

We revisited the BDEI model, which allows estimation of epidemiological parameters from genomic data for pathogens that feature an incubation period. We implemented a new maximum-likelihood BDEI parameter estimator, PyBDEI, which drastically increases parameter optimisation performance, accuracy and speed with respect to previously available estimators.

Previous implementations were either limited by data set size (in likelihood-based frameworks) or by heavy computational effort required to train the predictors, hence limiting the applicability outside of already pre-trained settings (in a machine learning framework): The Bayesian gold standard method BEAST2 required ~ 57 CPU hours for a 500-tip tree analysis (on a fixed input tree) and sometimes suffered from numerical instability; the maximum-likelihood implementation TreePar was limited to trees of ~ 50 tips; the deep-learning estimator PhyloDeep required hundreds of CPU hours for generating training data and training the predictors. However, the accuracy of estimations improves with the data set size (as expected, see our simulations). In the world of rapidly growing sequencing data sets (34), we can gain important insights on epidemic spreads by harvesting all available information.

Our fast and accurate maximum-likelihood estimator is applicable to very large data sets (2 minutes on a 10 000-tip tree), making parameter estimation instantaneous with respect to phylogenetic tree reconstruction times (hours). We obtained this performance by using the peculiarities of master equations behind the BDEI and general MTBD-family models, which permitted us to rewrite them in a branch-specific way. This representation features simple initial conditions with zero or one values, and avoids numerical problems that could occur in the original system due to very small positive values of the initial conditions. Even more importantly, our branch-specific representation removes the recursive dependency between ODEs corresponding to different tree nodes, and permits their time-consuming resolution to be performed in parallel and independently for each tree node. The results can then be combined into tree likelihood with a standard pruning algorithm. While the likelihood-combining step remains recursive, its time cost is negligible in comparison with the recursive ODE resolution used in the previous studies. Moreover, for cases where tree node states are known we obtained an explicit likelihood formula. It can be represented in a logarithmic form to avoid potential underflow issues during calculations. Tree node states could be known from metadata, or because they were generated by an MTBD process in which only one state can transmit or get sampled (e.g. BDEI). Our approach could be easily used in Bayesian setting as well, and could potentially be implemented in BEAST2.

Extension to forests permits the use of our estimator in situations where health policies change over time (providing a flexible alternative to the Bayesian skyline), as well as in situations of multiple (not necessarily simultaneous) pathogen introductions to a country of interest.

We applied our estimator to the 2014 Ebola epidemic in Sierra-Leone, after the introduction of quarantine measures. The analysis took < 2 hours (the majority of which was the tree reconstruction), and allowed us to estimate the reproductive number $R_e \approx 0.9$ suggesting contained epidemic, narrow down the incubation period estimate to 10–14 days, and estimate the sampling probability $\rho \approx 0.1$.

The results of our study break the computational bottleneck that was preventing phylodynamics from catching up with rapid pathogen genome sequencing. It opens way to fast and accurate estimations of epidemiological parameters for emerging and on-going epidemics.

Materials and Methods

Equivalence between Eq. [2] and Eq. [4]. The likelihood Eq. [2] for a tree $\mathscr{T}$ is recursive, and when using $P_I^{(i)}(t)$ and $P_E^{(i)}(t)$ needs to be resolved with a pruning algorithm while climbing the tree. However, we can transform it into a non-recursive Eq. [4] with $p_I^{(i)}(t)$ and $p_E^{(i)}(t)$, by alternating *replacement* and *unfolding* steps. A replacement step consists in replacing $P_s^{(i)}(t)$ ($s \in \{E, I\}$) with $p_s^{(i)}(t) \cdot C^{(i)}$ and is followed by an unfolding step. An unfolding step either (if i is a tip) unfolds $C^{(i)}$ into $\psi\rho$ and stops; or (if i is an internal node) unfolds $C^{(i)}$ into $\lambda\Big(P_I^{(l.c.\ of\ i)}(t_i)P_E^{(r.c.\ of\ i)}(t_i) + P_E^{(l.c.\ of\ i)}(t_i)P_I^{(r.c.\ of\ i)}(t_i)\Big)$, where l.c. and r.c. stand for left child and right child, and proceeds with replacements. In Eq. [5], we show the transformation process:

$$
\begin{aligned}
L(\mathscr{T}|\Theta) &= P_I^{(root)}(0) && \leftarrow \textit{Eq. [2]}\\
&= p_I^{(root)}(0) \cdot C^{(root)} && \leftarrow \textit{replacement}\\
&= 1 && \leftarrow p_I^{(root)}(0) = 1\\
&\quad \cdot \lambda\Big(P_I^{(l.c.\ of\ root)}(t_{root})P_E^{(r.c.\ of\ root)}(t_{root}) + P_E^{(l.c.\ of\ root)}(t_{root})P_I^{(r.c.\ of\ root)}(t_{root})\Big) && \leftarrow \textit{unfolding of } C^{(root)}\\
&= \lambda\Big(p_I^{(l.c.\ of\ root)}(t_{root})p_E^{(r.c.\ of\ root)}(t_{root}) + p_E^{(l.c.\ of\ root)}(t_{root})p_I^{(r.c.\ of\ root)}(t_{root})\Big) && \leftarrow \textit{replacement}\\
&\quad \cdot C^{(l.c.\ of\ root)} \cdot C^{(r.c.\ of\ root)}\\
&= \lambda\Big(p_I^{(l.c.\ of\ root)}(t_{root})p_E^{(r.c.\ of\ root)}(t_{root}) + p_E^{(l.c.\ of\ root)}(t_{root})p_I^{(r.c.\ of\ root)}(t_{root})\Big)\\
&\quad \cdot \begin{cases} \psi\rho, & \textit{if l.c. of root is a tip}\\ \lambda\Big(p_I^{(l.c.\ of\ l.c.\ of\ root)}(t_{l.c.\ of\ r.})p_E^{(r.c.\ of\ l.c.\ of\ root)}(t_{l.c.\ of\ r.}) + p_E^{(l.c.\ of\ l.c.\ of\ root)}(t_{l.c.\ of\ r.})p_I^{(r.c.\ of\ l.c.\ of\ root)}(t_{l.c.\ of\ r.})\Big) & \\ \cdot C^{(l.c.\ of\ l.c.\ of\ root)} \cdot C^{(r.c.\ of\ l.c.\ of\ root)}, & \textit{otherwise} \end{cases} && \leftarrow \textit{unfolding of } C^{(l.\ c.\ of\ root)}\\
&\quad \cdot \begin{cases} \psi\rho, & \textit{if r.c. of root is a tip}\\ \lambda\Big(p_I^{(l.c.\ of\ r.c.\ of\ root)}(t_{r.c.\ of\ r.})p_E^{(r.c.\ of\ r.c.\ of\ root)}(t_{r.c.\ of\ r.}) + p_E^{(l.c.\ of\ r.c.\ of\ root)}(t_{r.c.\ of\ r.})p_I^{(r.c.\ of\ r.c.\ of\ root)}(t_{r.c.\ of\ r.})\Big) & \\ \cdot C^{(l.c.\ of\ r.c.\ of\ root)} \cdot C^{(r.c.\ of\ r.c.\ of\ root)}, & \textit{otherwise} \end{cases} && \leftarrow \textit{unfolding of } C^{(r.\ c.\ of\ root)}\\
&= \dots && \leftarrow \textit{keep unfolding and replacing}\\
&= \prod_{i \in \textit{internal nodes}} \lambda\Big(p_I^{(l.c.\ of\ i)}(t_i)p_E^{(r.c.\ of\ i)}(t_i) + p_E^{(l.c.\ of\ i)}(t_i)p_I^{(r.c.\ of\ i)}(t_i)\Big) && \leftarrow \textit{unfolded and replaced } C^{(i)} \textit{ for } n-1 \textit{ internal nodes}\\
&\quad \cdot (\psi\rho)^n && \leftarrow \textit{unfolded } C^{(i)} \textit{ for } n \textit{ tips} \quad \leftarrow \textit{non-log version of Eq. [4]}
\end{aligned} \tag{5}
$$

Forest likelihood. Forest $\mathscr{F}$ likelihood [6] under the BDEI model with parameters $\Theta = \{\mu, \lambda, \psi, \rho\}$ generalizes tree likelihood Eq. [4] to the case of $f \geq 1$ observed and $u \geq 0$ hidden trees:

$$
\begin{aligned}
logL(\mathscr{F}, \Theta) = u log\Big(\pi_E U_E(0) + \pi_I U_I(0)\Big) & && \leftarrow \textit{unobserved trees}\\
+ n log(\psi\rho) & && \leftarrow \textit{sampling of } n \textit{ tips}\\
+ (n-f) log\lambda & && \leftarrow n-f \textit{ transmission events}\\
+ \sum_{j=1}^{f} log\Big(\pi_E p_E^{(root\ of\ forest\ j)}(0) + \pi_I p_I^{(root\ of\ forest\ j)}(0)\Big) & && \leftarrow f \textit{ root branch evolutions}\\
+ \sum_{i \in \textit{internal nodes}} log\Big(p_I^{(l.c.\ of\ i)}(t_i)p_E^{(r.c.\ of\ i)}(t_i) + p_E^{(l.c.\ of\ i)}(t_i)p_I^{(r.c.\ of\ i)}(t_i)\Big) & && \leftarrow \textit{child branch evolutions for } n-f \textit{ internal nodes}
\end{aligned} \tag{6}
$$

While in tree likelihood Eq. [4], we assumed that the epidemic started directly with the first transmission, for a forest we relax this assumption. The root of the tree in Fig. [1] is placed at $t = 0$, and does not have a branch (its length is zero). For trees in a forest we allow for non-zero root branches, which corresponds to their sub-epidemics starting some time before the first transmission. This implies that the states of the individuals represented by the root branches are unknown, and both I and E should be considered. We therefore combine the two possibilities weighting them by probabilities π_s of a root being in the corresponding state $s \in \{E, I\}$ at time $t = 0$. Assuming that the relative number of individuals in each state is at equilibrium, we can calculate π_E and π_I as described in the next subsection (Eq. 11).

In Eq. [6] we assumed that all the $f + u$ sub-epidemics in the forest $\mathscr{F}$ started at the same time ($t = 0$). This condition can be easily relaxed by replacing zeros with the corresponding starting times (for unobserved trees and for root branch evolutions).

Stationary distribution. Stationary state distribution $\Pi = \{\pi_E, \pi_I\} :\ \pi_E + \pi_I = 1$ corresponds to the ratios of states E and I at a given time t after these ratios stopped changing (assuming that this may happen). $\pi_s = \frac{N_s(t)}{N(t)}$, where $N_s(t)$ is the number of individuals of type $s \in \{E, I\}$ and $N(t)$ is the total number of infected (infectious or not) individuals at time t. Hence, the derivative of the number of individuals of type s is proportional to the derivative of the total number of infected individuals:

$$
dN_s(t) = \pi_s dN(t). \tag{7}
$$

The number of individuals in state I increases due to becoming infectious of individuals in state E and decreases due to removal, while the number of individuals in state E decreases due to becoming infectious and increases due to transmissions:

$$\begin{cases} dN_I(t) &= \mu N_E(t) - \psi N_I(t) \\ dN_E(t) &= -\mu N_E(t) + \lambda N_I(t) \end{cases}. \quad [8]$$

Becoming infectious changes the corresponding individual's state but does not affect the total number of infected individuals, transmissions increase the total number, and removal decreases it. Note that only individuals in state I can transmit or be removed:

$$dN(t) = \lambda N_I(t) - \psi N_I(t). \quad [9]$$

Combining [8] and [9] we rewrite [7] as a system of multivariate algebraic equations:

$$\begin{cases} \mu\pi_E - \psi\pi_I = \pi_I^2(\lambda - \psi) \\ \pi_E + \pi_I = 1 \end{cases}, \quad [10]$$

from which we derive a quadratic equation for π_I: $\pi_I^2(\lambda - \psi) + \pi_I(\mu + \psi) - \mu = 0$, and the following stationary distribution:

$$\begin{cases} \pi_I &= \begin{cases} \frac{\mu}{\mu+\psi}, & \text{if } \lambda = \psi \\ \frac{-(\mu+\psi)+\sqrt{(\mu-\psi)^2+4\mu\lambda}}{2(\lambda-\psi)}, & \text{otherwise} \end{cases} \\ \pi_E &= 1 - \pi_I \end{cases}. \quad [11]$$

PDF reconditioning and likelihood calculation for a general MTBD model. A general MTBD model describes m possible individual states, their state-change, removal and transmission rates, and a sampling probability upon removal ρ. An individual in state k can be removed at rate ψ_k, change their state to state l at rate μ_{kl} (where $\mu_{kk} = 0$), and transmit their pathogen to an individual in state l at rate λ_{kl}. The corresponding master equations are presented in System [12], which describes the probabilities $P_{kl}^{(i)}(t)$ of an individual evolving as observed in the tree, starting at time t in state k on a branch connecting a node i to its parent. The initial conditions are defined at time $t = t_i$ (i.e. at the node i). To account for incomplete sampling, the system also includes the probabilities $U_k(t)$ of evolving unobserved till the end of the sampling period (time T), starting at time t in state k.

$$\begin{cases} \dot{P}_{kl}^{(i)}(t) = & \left(\sum_{s=1}^{m} \mu_{ks} + \sum_{s=1}^{m} \lambda_{ks} + \psi_k\right) P_{kl}^{(i)}(t) \leftarrow \textit{no event in the next infinitesimal time } \Delta t \\ & - \sum_{s=1}^{m} \mu_{ks} P_{sl}^{(i)}(t) \leftarrow \textit{change of the state, followed by evolution from the new state} \\ & - \sum_{s=1}^{m} \lambda_{ks} P_{kl}^{(i)}(t) U_s(t) \leftarrow \textit{transmission, where the recipient subtree stayed unsampled} \\ & - \sum_{s=1}^{m} \lambda_{ks} P_{sl}^{(i)}(t) U_k(t) \leftarrow \textit{transmission, where the donor subtree stayed unsampled} \\ P_{kl}^{(i)}(t_i) = & C_{kl}^{(i)} = \begin{cases} 0, & \text{if } k \neq l \\ \psi_l \rho, & \text{if } i \text{ is a sampled tip in state } k = l \\ \sum_{s=1}^{m} \lambda_{ls} \left(P_l^{(l.c.\ of\ i)}(t_i) P_s^{(r.c.\ of\ i)}(t_i) + P_s^{(l.c.\ of\ i)}(t_i) P_l^{(r.c.\ of\ i)}(t_i) \right), & \text{if } i \text{ is an internal node in state } k = l \end{cases} \\ \dot{U}_k(t) = & \left(\sum_{s=1}^{m} \mu_{ks} + \sum_{s=1}^{m} \lambda_{ks} + \psi_k\right) U_k(t) \leftarrow \textit{no event in the next infinitesimal time } \Delta t \\ & - \sum_{s=1}^{m} \mu_{ks} U_s(t) \leftarrow \textit{change of the state, folowed by unsampled evolution from the new state} \\ & - \sum_{s=1}^{m} \lambda_{ks} U_k(t) U_s(t) \leftarrow \textit{transmission, followed by unsampled evolutions of both the donor and the recipient subtrees} \\ & -\psi_k(1-\rho) \leftarrow \textit{removal without sampling} \\ U_k(T) = & 1 \leftarrow \textit{the probability to stay unsampled over time 0 is 1} \end{cases} \quad [12]$$

Note that for a node i in state l the initial condition $C_{kl}^{(i)} = 0$ for all $k \neq l$, while $C_{ll}^{(i)} > 0$. This allows us to rescale the equations, using $p_{kl}^{(i)}(t) = P_{kl}^{(i)}(t)/C_{ll}^{(i)}$, in the same way as we did for the BDEI model. Conceptually, $p_{kl}^{(i)}(t)$ is a PDF of an individual evolving as on an observed branch that connects a node i to its parent, starting at time t in state k on this branch and finishing at time t_i in state l (at node i), without taking into account i's subtree and the event at node i. Like for the BDEI model, PDF reconditioning allows to (1) avoid potential underflow issues during equation resolution by having only zero and one initial condition values, and (2) perform costly numerical resolution of master equations for each tree node in parallel.

In the case where all node states are known (e.g. from metadata or due to model peculiarities, as for the BDEI case), using $p_{kl}^{(i)}(t)$ instead of $P_{kl}^{(i)}(t)$ also permits us to express tree likelihood for model parameters Θ in a non-recursive way, and easily transform it to a logarithmic form (to avoid underflow issues while multiplying small numbers):

$$\begin{aligned} logL(\mathscr{T}|\Theta) = & \sum_{i \in tips} log(\psi_{state(i)}\rho) && \leftarrow \textit{sampling of } n \textit{ tips} \\ & + \sum_{i \in internal\ nodes} log\Bigg(\sum_{k=1}^{m} \lambda_{state(i),k} \cdot && \leftarrow n-1 \textit{ transmission events} \\ & \cdot \Big(p_{state(i),state(l.c.\ of\ i)}^{(l.c.\ of\ i)}(t_i) p_{k,state(r.c.\ of\ i)}^{(r.c.\ of\ i)}(t_i) && \leftarrow \textit{child branch evolutions} \\ & + p_{k,state(l.c.\ of\ i)}^{(l.c.\ of\ i)}(t_i) p_{state(i),state(r.c.\ of\ i)}^{(r.c.\ of\ i)}(t_i) \Big) \Bigg) && \textit{for each internal node} \end{aligned} \quad [13]$$

However, for the general case, the combination of different internal node state configurations into tree likelihood formula need to be performed with a pruning algorithm (18). The likelihood-combining tree traversal starts from the tips and climbs the tree till the root, while calculating a subtree likelihood $L_k^{(i)}(\Theta)$ for each visited node i for each possible state k:

$$L_k^{(i)}(\Theta) = \begin{cases} 0, & \textit{if } i\textit{'s state is not } k \\ \psi_k \rho, & \textit{if } i \textit{ is a tip whose state can be } k \\ \sum_{l=1}^{m} \lambda_{kl} \cdot \\ \quad \cdot \Big(\sum_{s=1}^{m} p_{ks}^{(l.c.\ of\ i)}(t_i) L_s^{(l.c.\ of\ i)}(\Theta) \cdot \sum_{s=1}^{m} p_{ls}^{(r.c.\ of\ i)}(t_i) L_s^{(r.c.\ of\ i)}(\Theta) \\ \quad + \sum_{s=1}^{m} p_{ks}^{(r.c.\ of\ i)}(t_i) L_s^{(r.c.\ of\ i)}(\Theta) \cdot \sum_{s=1}^{m} p_{ls}^{(l.c.\ of\ i)}(t_i) L_s^{(l.c.\ of\ i)}(\Theta) \Big), & \textit{if } i \textit{ is an internal node whose state can be } k \end{cases} \quad [14]$$

Tree likelihood then can be calculated as the root likelihood:

$$L(\mathscr{T}|\Theta) = \sum_{k=1}^{m} \pi_i L_s^{(root)}(\Theta), \textit{ where } \pi_i \textit{ is the equilibrium frequency of state } k\textit{, derived in (22).} \quad [15]$$

Note that unlike the known-tree-node-state likelihood (Eq. [13]), the recursive unknown-tree-node-state likelihood (Eq. [15]) does not allow for an easy logarithmic representation, and hence is prone to underflow issues (as the original MTBD representation). Its calculation on large trees therefore requires additional small number rescaling techniques as recently proposed in (16) and very common in phylogenetic inference.

Tree reconstruction for Ebola SLE epidemic analysis. We reconstructed a maximum-likelihood phylogeny of 1 610 tips for the Ebola samples from (33) with RAxML-NG (v1.0.2, GTR+G4+FO+IO) (35), and rooted it based on sampling dates using LSD2 (v1.10) (36). As Ebola's mutation rate is slower than its transmission rate, the initial phylogeny contained 246 polytomies (i.e. multiple transmissions, which happened faster than the virus acquired a mutation, hence making them undistinguishable in the phylogeny). The BDEI model, on the other hand, assumes a binary tree. We therefore resolved these polytomies randomly (10 times, to check for robustness of the estimates) using a coalescent approach.

We then dated each of the 10 trees with LSD2 (36) (v1.10: github.com/tothuhien/lsd2/tree/v1.10, under strict molecular clock with outlier removal) using tip sampling dates, and reconstructed the ancestral characters for country with PastML (37) (v1.9.34, MPPA+F81).

Lastly, we extracted 10 SLE forests from these trees to represent the Ebola epidemic in SLE between July 30 2014 (when the SLE government began to deploy troops to enforce quarantines according to news24.com) and September 12 2015 (the last SLE sample in our dataset) by (1) cutting each tree on July 30 2014 to remove the more ancient part (with a different health policy); (2) among the July-31-on trees, picking those whose root's predicted character state for country was SLE (light-green branches at the level of July 31 2014 in Fig. S2); (3) removing the non-SLE subtrees (indicated with other colours in Fig. S2) from the selected July-31-on SLE trees to focus on the epidemic within the country, without further reintroductions.

The reconstruction took 1 hour for the phylogeny, 10 minutes for tree dating, and 1 minute for country ancestral character prediction.

BEAST2 and PhyloDeep settings. BEAST2 (v2.6.2 with package bdmm (16) v1.0) was configured for 10^6 MCMC steps with the following priors: $\mu \in U(0.02, 5.0)$, $\psi \in U(0.1, 1.0)$, and ρ fixed to the real value. The initial values in the MCMC were set to the medians observed in the PhyloDeep training set, namely $\mu = 2.51$, $\psi = 0.55$, and $R_0 = 3$. The tree was fixed to the real tree. For each tree, the Effective Sample Size (ESS) on all parameters was evaluated, and the median of a posteriori values was reported, corresponding to all recorded steps (i.e. actual MCMC steps spaced by 1 000) past the 10% burn-in. For simulations for which BEAST2 did not converge (2%) after 10^6 MCMC steps, the median of the parameter distribution used for tree simulations was reported instead.

PhyloDeep (v0.2.51) was run with ρ fixed to the real value and Convolutional Neural Networks trained on the Compact Bijective Ladderized Vector full tree representation (CNN-CBLV).

The visualisations of the analyses of simulated data were performed with the Python 3 library seaborn (42, 43).

Code and data availability. The core of our estimator is implemented in C++ and uses the NLopt library (38) for non-linear optimization. The parallelization is achieved with the C++ `thread_pool` tools (39). To facilitate the use of our estimator in Python and perform additional validation of input trees, we wrapped the core estimator into a Python 3 library PyBDEI. PyBDEI uses ETE 3 framework for tree manipulation (40) and NumPy package for array operations (41).

Our estimator is available as a command-line program and a Python 3 library via PyPi (pybdei), and via Docker/Singularity (evolbioinfo/bdei). Its source code, the simulated and real data used for its assessment, as well as the Snakemake (44) data analysis pipelines, and the installation and usage documentation are available on GitHub at github.com/evolbioinfo/bdei. BEAST2 xml files and command lines are available at github.com/evolbioinfo/phylodeep/tree/main/data_publication.

ACKNOWLEDGMENTS. The authors would like to thank Dr Jakub Voznica for fruitful discussions on different BDEI implementations.

1. BT Grenfell, et al., Unifying the Epidemiological and Evolutionary Dynamics of Pathogens. Science **303** (2004).
2. EM Volz, K Koelle, T Bedford, T Bhattacharya, E Delaporte, Viral Phylodynamics. PLoS Comput. Biol. **9**, e1002947 (2013).
3. EM Volz, SL Kosakovsky Pond, MJ Ward, AJ Leigh Brown, SD Frost, Phylodynamics of infectious disease epidemics. Genetics **183**, 1421–1430 (2009).
4. AJ Drummond, A Rambaut, B Shapiro, OG Pybus, Bayesian coalescent inference of past population dynamics from molecular sequences. Mol. Biol. Evol. **22**, 1185–1192 (2005).
5. OG Pybus, A Rambaut, PH Harvey, An integrated framework for the inference of viral population history from reconstructed genealogies. Genetics **155**, 1429–1437 (2000).
6. DG Kendall, On the generalized "birth-and-death" process. Ann. Math. **19**, 1–15 (1948).
7. W Maddison, P Midford, S Otto, Estimating a Binary Character's Effect on Speciation and Extinction. Syst. Biol. **56**, 701–710 (2007).
8. T Stadler, On incomplete sampling under birth-death models and connections to the sampling-based coalescent. J. Theor. Biol. **261**, 58–66 (2009).
9. T Stadler, Sampling-through-time in birth-death trees. J. Theor. Biol. **267**, 396–404 (2010).
10. A Macpherson, S Louca, A Mclaughlin, JB Joy, MW Pennell, Unifying Phylogenetic Birth-Death Models in Epidemiology and Macroevolution. Syst. Biol. **71** (2021).
11. T Stadler, S Bonhoeffer, Uncovering epidemiological dynamics in heterogeneous host populations using phylogenetic methods. Philos. Transactions Royal Soc. B: Biol. Sci. **368**, 20120198–20120198 (2013).
12. T Stadler, D Kühnert, DA Rasmussen, L du Plessis, Insights into the Early Epidemic Spread of Ebola in Sierra Leone Provided by Viral Sequence Data. PLoS Curr. **6** (2014).
13. HW Hethcote, The mathematics of infectious diseases. SIAM review **42**, 599–653 (2000).
14. AJ Drummond, A Rambaut, BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol. Biol. **7**, 214 (2007).
15. R Bouckaert, et al., BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. PLoS Comput. Biol. **15** (2019).
16. J Scire, J Barido-Sottani, D Kühnert, TG Vaughan, T Stadler, Improved multi-type birth-death phylodynamic inference in BEAST 2. bioRxiv , 2020.01.06.895532 (2020).
17. J Voznica, et al., Deep learning from phylogenies to uncover the transmission dynamics of epidemics. Nat. Commun. **13**, 3896 (2022).
18. J Felsenstein, Maximum Likelihood and Minimum-Steps Methods for Estimating Evolutionary Trees from Data on Discrete Characters. Syst. Biol. **22**, 240–249 (1973).

19. SA Berger, A Stamatakis, Accuracy and Performance of Single versus Double Precision Arithmetics for Maximum Likelihood Phylogeny Reconstruction. Lect. Notes Comput. Sci. (including subseries Lect. Notes Artif. Intell. Lect. Notes Bioinformatics) **6068 LNCS**, 270–279 (2009).
20. D Defour, Accuracy of a Maximum Likelihood Phylogeny Reconstruction, Technical report (2010).
21. A Zhukova, et al., Origin, evolution and global spread of SARS-CoV-2. Comptes Rendus. Biol. **0**, 1–20 (2020).
22. T Stadler, D Kühnert, S Bonhoeffer, AJ Drummond, Birth-death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV). Proc. Natl. Acad. Sci. United States Am. **110**, 228–33 (2013).
23. D Kühnert, T Stadler, TG Vaughan, AJ Drummond, Phylodynamics with Migration: A Computational Framework to Quantify Population Structure from Genomic Data. Mol. biology evolution **33**, 2102–16 (2016).
24. K Svanberg, A class of globally convergent optimization methods based on conservative convex separable approximations. SIAM J. on Optim. **12**, 555–573 (2002).
25. J Crank, P Nicolson, A practical method for numerical evaluation of solutions of partial differential equations of the heat-conduction type. Math. Proc. Camb. Philos. Soc. **43**, 50–67 (1947).
26. JC Butcher, Numerical methods for ordinary differential equations eds. C Vuik, P van Beek, F Vermolen, J van Kan. (Wiley), 3 edition, p. 540 (2016).
27. SS Wilks, The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses. The Annals Math. Stat. **9**, 60–62 (1938).
28. WER Team, Ebola Virus Disease in West Africa — The First 9 Months of the Epidemic and Forward Projections. N Engl J Med **371**, 1481–1495 (2014).
29. WER Team, West African Ebola Epidemic after One Year — Slowing but Not Yet under Control. N Engl J Med **372**, 584–587 (2015).
30. CM Rivers, ET Lofgren, M Marathe, S Eubank, BL Lewis, Modeling the Impact of Interventions on an Epidemic of Ebola in Sierra Leone and Liberia. PLoS Curr. **6** (2014).
31. MD Van Kerkhove, AI Bento, HL Mills, NM Ferguson, CA Donnelly, A review of epidemiological parameters from Ebola outbreaks to inform early public health decision-making. Sci. Data **2**, 1–10 (2015).
32. SK Gire, et al., Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. Science **345**, 1369–1372 (2014).
33. G Dudas, et al., Virus genomes reveal factors that spread and sustained the Ebola epidemic. Nature **544**, 309–315 (2017).
34. EB Hodcroft, et al., Want to track pandemic variants faster? Fix the bioinformatics bottleneck. Nature **591**, 30–33 (2021).
35. AM Kozlov, D Darriba, T Flouri, B Morel, A Stamatakis, RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. Bioinformatics **35**, 4453–4455 (2019).
36. TH To, M Jung, S Lycett, O Gascuel, Fast Dating Using Least-Squares Criteria and Algorithms. Syst. biology **65**, 82–97 (2016).
37. SA Ishikawa, A Zhukova, W Iwasaki, O Gascuel, A fast likelihood method to reconstruct and visualize ancestral scenarios. Mol. biology evolution **36**, 2069–2085 (2019).
38. SG Johnson, The nlopt nonlinear-optimization package ( ) Accessed: 2021-01-26.
39. A Williams, C++ concurrency in action: practical multithreading; 1st ed. (Manning Publ., Shelter Island, NY), (2012).
40. J Huerta-Cepas, F Serra, P Bork, ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. Mol. Biol. Evol. **33**, 1635–1638 (2016).
41. CR Harris, et al., Array programming with NumPy. Nature **585**, 357–362 (2020) Publisher: Nature Publishing Group.
42. ML Waskom, seaborn: statistical data visualization. J. Open Source Softw. **6**, 3021 (2021).
43. JD Hunter, Matplotlib: A 2D graphics environment. Comput. Sci. & Eng. **9**, 90–95 (2007).
44. J Köster, S Rahmann, Snakemake-a scalable bioinformatics workflow engine. Bioinformatics **28**, 2520–2522 (2012).

Supporting Information Appendix (SI).

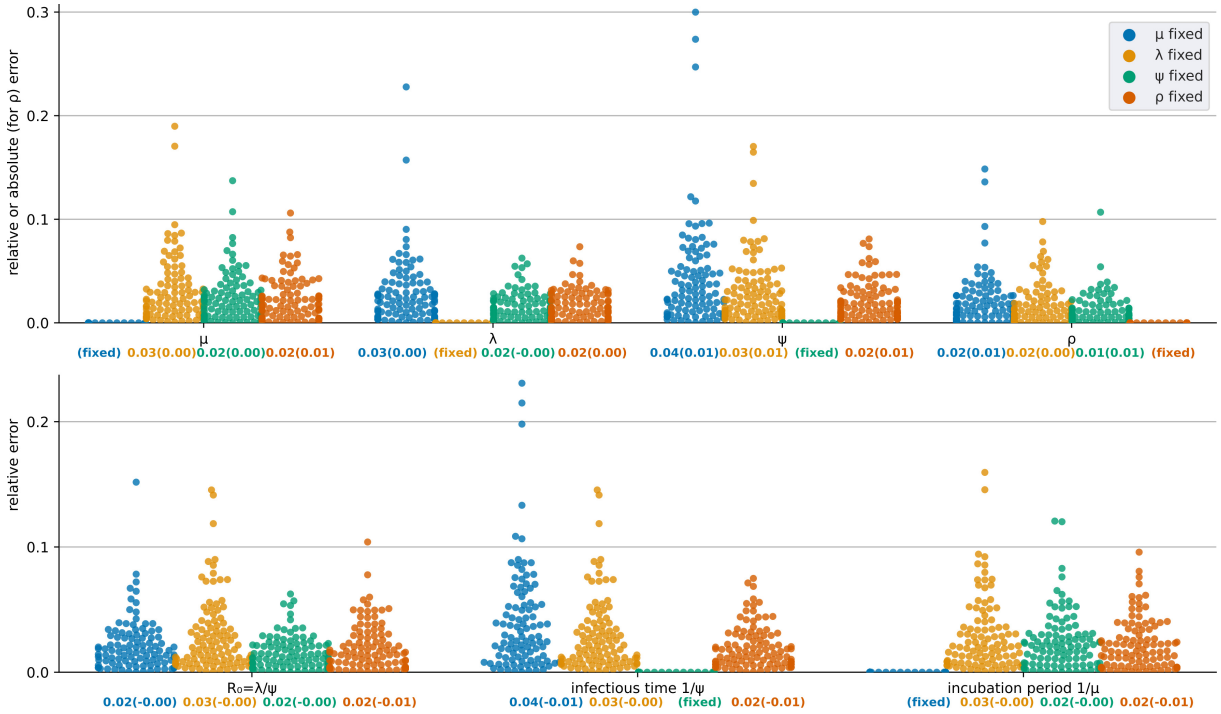

Fig. S 1. Comparison of inference accuracy on 100 test trees of 500–10 000 tips from (17), with (blue) μ fixed to the real value, (yellow) λ fixed to the real value, (green) ψ fixed to the real value, (orange) ρ fixed to the real value. We show the swarmplots of relative errors for each test tree and parameter. Average relative error (and in parentheses average relative bias) are displayed for each parameter and setting below their swarmplot. For ρ, absolute errors and bias are shown instead of the relative ones.

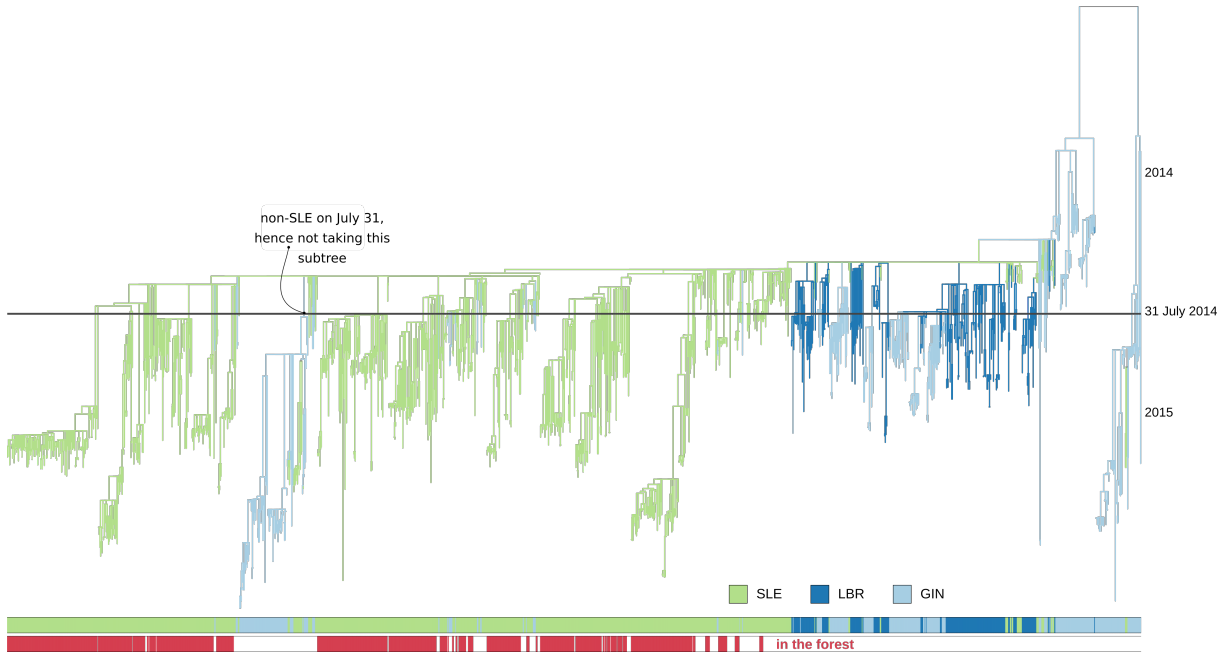

Fig. S 2. Ebola $2014-2016$ epidemic timetree (data from (33), polytomies are resolved as in forest 1) coloured by country predicted by PastML (37): Guinea (GIN) is light-blue, Liberia (LBR) is dark-blue, Sierra-Leone (SLE) is light-green. The bottom colourstrip shows the samples kept for the SLE 2014 epidemic analysis (forest 1): SLE samples that are directly related to the SLE epidemic of 31 July 2014 (start of quarantine) and sampled after this date. The SLE samples that were later reintroduced to SLE via other countries (e.g. in the indicated GIN-rooted subtree) were not included in the analysis.

Table S 1. Parameters estimated for SLE 2014 Ebola epidemic.

forest	tips	trees	hidden trees	infectious period $\frac{1}{\psi}$ (fixed)	R_e	incubation period $\frac{1}{\mu}$	ρ
1	779	56	477	2.6	0.97 [0.92 − 1.03]	14.3 [12.5 − 16.3]	0.13 [0.11 − 0.16]
1	779	56	477	5.0	0.98 [0.92 − 1.04]	13.9 [12.1 − 16.1]	0.16 [0.13 − 0.18]
1	779	56	744	2.6	0.95 [0.91 − 0.99]	12.1 [10.5 − 13.9]	0.09 [0.07 − 0.11]
1	779	56	744	5.0	0.95 [0.90 − 1.00]	11.0 [9.6 − 13.1]	0.11 [0.09 − 0.13]
2	762	68	465	2.6	0.97 [0.92 − 1.02]	13.8 [12.2 − 15.8]	0.12 [0.10 − 0.15]
2	762	68	465	5.0	0.97 [0.91 − 1.02]	13.5 [11.7 − 15.5]	0.15 [0.13 − 0.17]
2	762	68	732	2.6	0.95 [0.91 − 0.99]	11.7 [10.0 − 13.4]	0.09 [0.07 − 0.11]
2	762	68	732	5.0	0.94 [0.90 − 0.99]	10.7 [9.0 − 12.6]	0.10 [0.08 − 0.12]
3	764	70	463	2.6	0.97 [0.92 − 1.02]	14.3 [12.5 − 16.2]	0.13 [0.11 − 0.15]
3	764	70	463	5.0	0.97 [0.91 − 1.03]	13.8 [12.2 − 16.0]	0.15 [0.13 − 0.18]
3	764	70	730	2.6	0.95 [0.91 − 0.99]	12.0 [10.4 − 13.8]	0.09 [0.07 − 0.11]
3	764	70	730	5.0	0.94 [0.90 − 0.99]	11.0 [9.4 − 13.0]	0.10 [0.08 − 0.13]
4	759	63	470	2.6	0.97 [0.92 − 1.02]	14.2 [12.3 − 16.2]	0.12 [0.10 − 0.15]
4	759	63	470	5.0	0.97 [0.92 − 1.03]	13.7 [12.1 − 15.9]	0.15 [0.12 − 0.18]
4	759	63	737	2.6	0.95 [0.91 − 0.99]	12.0 [10.3 − 13.8]	0.09 [0.07 − 0.11]
4	759	63	737	5.0	0.95 [0.90 − 0.99]	10.9 [9.1 − 12.9]	0.10 [0.08 − 0.12]
5	711	55	478	2.6	0.97 [0.92 − 1.02]	13.0 [11.3 − 14.9]	0.12 [0.10 − 0.14]
5	711	55	478	5.0	0.97 [0.91 − 1.02]	12.6 [10.6 − 14.7]	0.14 [0.11 − 0.17]
5	711	55	745	2.6	0.95 [0.91 − 0.99]	10.9 [9.2 − 12.7]	0.08 [0.06 − 0.10]
5	711	55	745	5.0	0.94 [0.90 − 0.99]	9.7 [7.9 − 11.3]	0.09 [0.07 − 0.11]
6	770	62	471	2.6	0.97 [0.92 − 1.02]	14.1 [12.3 − 16.0]	0.13 [0.11 − 0.16]
6	770	62	471	5.0	0.97 [0.92 − 1.03]	13.8 [11.8 − 15.8]	0.15 [0.12 − 0.19]
6	770	62	738	2.6	0.95 [0.91 − 0.99]	11.9 [10.3 − 13.7]	0.09 [0.07 − 0.11]
6	770	62	738	5.0	0.95 [0.90 − 0.99]	10.9 [9.0 − 12.8]	0.10 [0.08 − 0.12]
7	765	65	468	2.6	0.97 [0.92 − 1.02]	13.6 [11.9 − 15.5]	0.12 [0.10 − 0.15]
7	765	65	468	5.0	0.97 [0.92 − 1.03]	13.2 [11.2 − 15.3]	0.15 [0.12 − 0.17]
7	765	65	735	2.6	0.95 [0.91 − 0.99]	11.5 [9.9 − 13.3]	0.09 [0.07 − 0.10]
7	765	65	735	5.0	0.95 [0.90 − 0.99]	10.4 [8.6 − 12.0]	0.10 [0.08 − 0.12]
8	759	64	469	2.6	0.97 [0.92 − 1.02]	14.4 [12.5 − 16.5]	0.12 [0.10 − 0.15]
8	759	64	469	5.0	0.97 [0.92 − 1.03]	13.9 [11.8 − 16.1]	0.15 [0.12 − 0.17]
8	759	64	736	2.6	0.95 [0.91 − 0.99]	12.1 [10.6 − 13.8]	0.09 [0.07 − 0.10]
8	759	64	736	5.0	0.95 [0.90 − 0.99]	11.0 [9.2 − 13.1]	0.10 [0.08 − 0.12]
9	755	68	465	2.6	0.97 [0.92 − 1.02]	13.3 [11.8 − 15.1]	0.11 [0.10 − 0.14]
9	755	68	465	5.0	0.97 [0.91 − 1.02]	12.8 [10.9 − 14.9]	0.14 [0.11 − 0.16]
9	755	68	732	2.6	0.95 [0.91 − 0.99]	11.2 [9.6 − 12.9]	0.08 [0.06 − 0.10]
9	755	68	732	5.0	0.95 [0.90 − 0.99]	10.1 [8.3 − 11.9]	0.09 [0.07 − 0.11]
10	759	70	463	2.6	0.97 [0.92 − 1.02]	13.9 [12.1 − 15.9]	0.12 [0.10 − 0.15]
10	759	70	463	5.0	0.97 [0.91 − 1.02]	13.7 [11.7 − 15.7]	0.15 [0.12 − 0.17]
10	759	70	730	2.6	0.95 [0.91 − 0.99]	11.7 [10.0 − 13.6]	0.09 [0.07 − 0.10]
10	759	70	730	5.0	0.94 [0.90 − 0.99]	10.8 [9.2 − 12.7]	0.10 [0.08 − 0.12]