IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Digital Object Identifier

Bayesian Local Phylodynamics for Outbreak Detection

David Helekal, Xavier Didelot

ABSTRACT In epidemiology, it is often desirable to be able to reconstruct the history of a pathogen population and it's structure. The problem of reconstructing the history of a pathogen population can be approached using phylodynamics. Phylodynamics utilises genomic data to assemble phylogenies, which are then used to infer the population size history. This is possible by viewing a phylogeny as a realisation of a coalescent process, with appropriately rescaled time. This approximation can be justified by viewing the coalescent as a Moran model, run backwards in time with the time rate equal to the population size. Within this paper we will first introduce the coalescent process for phylodynamic inference, review it's inhomogeneous generalisation, and finally introduce the main result of this work, a new model capable of performing local phylodynamic inference, i.e. different inferences about the population hsitory on subsets of the whole population capable of detecting and modeling clonal expansions.

I. INTRODUCTION

CLONAL expansions are a process in which a particular subset of a given bacterial strain undergoes explosive population growth that can be traced back to a particular individual [1]. The presence of clonal expansions in bacterial populations have been of long-standing interest and is implicated in epidemic processes, were an outbreak can be traced to a single ancestor [1, 2, 3, 4]. This often happens when a particular strain or individual obtains a variant of a particular gene that confers evolutionary advantage, for example, antibiotic resistance [5, 6, 4].

The presence of clonal expansions leaves an imprint in the overall population structure of a given bacterial strain, wotj the particular topology associated with this often being referred to as star-like [1, 2]. The problem of detecting hidden population structure corresponding to clonal expansions has become a problem of interest in epidemiology and outbreak surveillance [7].

While methods for detecting inhomogeneities in the population structure and size have been of interest since the early days of genetic sequencing [1, 2], the interest in the problem increased with whole genome sequencing becoming more accessible and affordable [5, 8, 9].

Despite the problems of inferring population size from a genealogy and detecting heterogeneities in the population size of the entire population being intrinsically tied, all but a couple of methods [7, 10], to our knowledge, rely either on manual detection or indirect detection. We aim to propose a model for the formation of clonal expansions in genealogies using a structured coalescent process, and devise a bayesian method for joint estimation and detection of relative population sizes and clonal expansions.

II. EXISTING WORK

A. PHYLODYNAMIC METHODS

A phylogeny is a labeled tree, where leaf nodes correspond to sampling events, and internal nodes to the divergence of two separate clades from a common ancestor. Branch length labels correspond to time. As such a phylogeny can be viewed as a realisation of Kingman's Coalescent process which will now be briefly introduced, and the justification behind why viewing a phylogeny as its realisation outlined.

Kingman's Coalescent is a continuous time markov chain stochastic process, defined on the state space $1, 2, 3, ..., K$ which can be interpreted as a set of j particles, where each pair of particles independently coalesces at a constant rate. This gives transition rates:

$$g(j, j-1) = \binom{j}{2}\lambda \quad \lambda \in \mathbb{R}^+ \tag{1}$$

[11]

By taking a backwards in time approximation of the Moran model, the coalescent process can be modified to model evolution of genealogies, i.e. how do the ancestors

of a set of individuals relate to each other backwards in time. Denote the relative population size at time t by $\alpha(t)$. Under such modification the transition rates become:

$$g(j, j-1) = \binom{j}{2} \cdot \frac{1}{\alpha(t)} \tag{2}$$

[12]
One way to interpret this is as a rescaling of time inversely proportional to the population size under Wright-Fisher model [13].
As the transition rates depend on the relative population size, it is possible to utilise this model for the inverse problem of determining the history of the size of a population, based on genealogies reconstructed from genomic samples. Such methods are often referred to as SkyGrid methods, have been first introduced in [14] and [15].
The framework has then been extended to allow for piecewise continuous population size functions, referred to as SkyGrid [16] and to include covariates [17].

B. LOCAL PHYLODYNAMICS

Whereas phylodynamics considers the size history of an entire population, the idea behind local phylodynamics is to consider the histories of selected subsets of related individuals, where each subset can be traced to a single ancestor. This can for example be practical when trying to detect hidden outbreaks, or clonal expansions due to a particular strain of a bacteria gaining a fitness advantage. This idea has been investigated in [7], where a null hypothesis based testing framework is developed to identify subsets of individuals that seem to have a significantly different population history than the the rest of the phylogeny. Recently, a birth-death type model that allows for heterogeneous growth rate parameters has been introduced in [10].

III. METHODS

A. COALESCENT PRELIMINARIES

The coalescent process can be characterised as a time-inhomogeneous pure-death markov process. We now conduct a brief analysis of the process and re-derive some properties.
The waiting times can be derived as follows. For an inhomogenous CTMC, let $E_j(t)$ be the total exit rate from state j at time t. By the markov property individual exit events from a given state only depend on the state and given time, i.e. they form a time-inhomogeneous poisson process. As such the probability of no events in an interval $[t, t+s] \quad s \in \mathbb{R}^+$ is

$$\begin{aligned} &\exp\left(-\int_t^{t+s} E_j(\tau)d\tau\right) = \\ &\exp\left(-\int_0^{s} E_j(t+\tau)d\tau\right) \end{aligned} \tag{3}$$

The waiting times are defined as

$$W_j(t) = \inf\{s : X(t+s) \neq j \mid X(t) = j\} \tag{4}$$

As such

$$W_j(t) > s \Rightarrow \forall \tau \in [t, t+s] \quad X(\tau) = j \tag{5}$$

Furthermore the above relation holds iff no exit event have occurred in the time interval $[t, t+s]$. As such:

$$\begin{aligned} P[W_j(t) > s] \quad &= P[\text{no exit events in } [t, t+s]] \\ &= \exp\left(-\int_0^{s} E_j(t+\tau)d\tau\right) \end{aligned} \tag{6}$$

And as such:

$$P[W_j(t) < s] = 1 - \exp\left(-\int_0^{s} E_j(t+\tau)d\tau\right) \tag{7}$$

In the case of phylodynamic coalescent this becomes

$$P[W_j(t) \leq s] = 1 - \exp\left(-\int_0^{s} \frac{\binom{j}{2}}{\alpha(t+\tau)}d\tau\right) \tag{8}$$

From equation 8 we can see that the waiting times between individual coalescent events depend on the relative population size.

B. SIMULATING PHYLOGENIES

We are interested in simulating the scenario where sampling events and the relative population size is given. Denote the collection of sampling events in decreasing time order by:

$$\{s_j\}_{1 \leq j \leq K}, \quad i > l \Leftrightarrow s_i \leq s_l \quad \forall i, l \tag{9}$$

Under this scenario, phylogenies can be simulated as realisations of the inhomogenous coalescent process, conditioned on sampling events, using modified Gillespie's Algorithm [18]. The key difference between standard Gillespie's Algorithm setting and our problem setting is that we need to condition on the sampling events.
Hence at any time greater than the earliest sampling event $t_i > s_K$ in the simulation, we first need to determine whether an event happens or not in the interval $[s_j, t_i]$ with $s_j = max\{s \in S : s < t_i\}$. Define $\Delta t_i \triangleq t_i - s_j$, and assume there are j individual clades extant at time t. We are interested in the probability

$$P[W_j(t_i) > \Delta t_i] = \exp\left(-\int_0^{\Delta t_i} \frac{\binom{j}{2}}{\alpha(t+\tau)}d\tau\right)$$

To determine the next in simulation draw $r \sim \mathtt{U}([0,1])$. If $r < P[W_j(t_i) > \Delta t_i]$, no events happen in the interval $[s_j, t_i]$. As such set $t_{i+1} = s_j$ and repeat the process.
If $r > P[W_j(t_i) > \Delta t_i]$ a coalescent event must happen in the interval $[s_j, t_i]$. Therefore we need to determine

the next waiting time $W_j(t_i)$. The cumulative distribution function (CDF) of $W_j(t_i)$ is given by

$$P[W_j(t_i) < c \mid W_j(t_i) < \Delta t_i] = \frac{1 - \exp\left(-\int_0^c \frac{\binom{j}{2}}{\alpha(t+\tau)} d\tau\right)}{1 - P[W_j(t_i) > \Delta t_i]} \tag{10}$$

$W_j(t_i)$ can be generated by the inverse transform of exponentially distributed variables.
Let $W \sim \texttt{exp}(1)$ and note that random variates W' with the CDF

$$P[W' \leq c] = P[W \leq c \mid W \leq \Delta' t] \quad \forall c \leq \Delta t \tag{11}$$

can be generated from random variables $u \sim \texttt{U}([0,1])$ by the following inverse transform

$$T(u) = -1 \log[1 - u(1 - \exp(-\Delta' t))] \tag{12}$$

This can be justified as follows

$$P[T(u) \leq c] = P[u \leq T^{-1}(c)]$$

Using the property that $P[u \leq a] = a$ and requiring that $T(u)$ follows the same distribution as W' we obtain

$$\Rightarrow \quad P[u \leq T^{-1}(c)] = \frac{\int_0^c \exp(-t)dt}{\int_0^{\Delta t} \exp(-t)dt}$$

$$\Rightarrow \quad T^{-1}(c) = \frac{1 - \exp(-c)}{1 - \exp(-\Delta t)}$$

Being abled to generate appropriate random variates W', a further inverse transform is applied, this time to obtain $W_j(t)$ from W'. First, we require that $\Delta' t$ be such that $P[W < \Delta' t]$ is equal to $P[W_j(t_i) < \Delta t]$. Next, we seek a function $G(W'; t)$ such that

$$P[W' \leq G^{-1}(c;t)] = P[W_j(t) \leq c \mid W_j(t) \leq \Delta t]$$

$$\Rightarrow \quad \frac{1 - \exp\left(-G^{-1}(c;t)\right)}{1 - P[W > \Delta' t]} = \frac{1 - \exp\left(-\int_0^c \frac{\binom{j}{2}}{\alpha(t+\tau)} d\tau\right)}{1 - P[W_j(t_i) > \Delta t_i]} \tag{13}$$

Using that $1 - P[W > \Delta' t] = 1 - P[W_j(t_i) > \Delta t_i]$ we obtain

$$\Rightarrow \quad 1 - \exp\left(-G^{-1}(c;t)\right) = 1 - \exp\left(-\int_0^c \frac{\binom{j}{2}}{\alpha(t+\tau)} d\tau\right) \tag{14}$$

Where the inverse transform G^{-1} is given by

$$\Rightarrow G^{-1}(c;t) = \int_0^c \frac{\binom{j}{2}}{\alpha(t+\tau)} d\tau \tag{15}$$

Therefore the scheme to generate waiting times $W_j(t)$ is to first generate appropriate W' from uniformly distributed random variates, and then to further transform W' by solving 15.
In practice, 15 often cannot be solved analytically, and hence we resort to numeric methods.

C. INFERENCE UNDER INHOMOGENOUS COALESCENT

The derivation of the likelihood function for time-inhomogenous coalescent process can be found in [15]. We adapt notational convention from [15]. Let $\{t_i\}_{i \in S \subset \mathbb{N}}$ denote the times of events in increasing order. Let $t_Y \triangleq \{t_i\}_{i \in Y}$ denote times of coalescent events and $t_I \triangleq \{t_i\}_{i \in I}$ denote times of sampling events, where Y, I are disjoint partitions of the index set S with the property that $S = Y \cup I$. The likelihood of a particular genealogy is then given by:

$$\mathcal{L}(g \mid \alpha) = \prod_{i \in S \setminus 1} \left(\mathbb{1}_Y(i) \frac{\binom{k_i}{2}}{\alpha(t_i)} + \mathbb{1}_I(i)\right) \times \exp\left(-\int_{t_{i-1}}^{t_i} \frac{\binom{k_i}{2}}{\alpha(\tau)} d\tau\right) \tag{16}$$

The log-likelihood is:

$$\log \mathcal{L}(g \mid \alpha) = -\sum_{i \in S \setminus 1} \int_{t_{i-1}}^{t_i} \frac{\binom{k_i}{2}}{\alpha(\tau) d\tau} + \sum_{i \in Y} \log \frac{\binom{k_i}{2}}{\alpha(t_i)} \tag{17}$$

D. COALESCENT WITH LOCAL POPULATION STRUCTURE

We now present the novel local phylodynamic model that we developed. In order to help illustrate the concepts behind this model, consider the following scenario. Suppose we sequence the genomes of a set of pathogenic bacterial samples. At an unknown point in time a particular strain acquired a mutation which conferred resistance to a widely used antibiotic. This increases the particular strains fitness and enables it to undergo a period of rapid growth leading to a clonal expansion. Assuming that this increase in fitness occurs in a short time span, the clade – a set of lineages sharing the same common ancestor, of this strain will behave differently in the phylogenetic tree having a coalescent rate corresponding to a rapidly expanding population starting from a very small number of individuals.
The problem of identifying hidden population structure has been proposed in [7], where a testing based approach was used to identify structure in a phylogeny, as well as in [10], where a birth-death type model is used.
In our approach, we will build upon the standard coalescent model, modifying it as to allow for change points located on the branches of the phylogeny, marking the event when a particular clade starts behaving according to a different population size function than its parent clade.
In our model, coalescent nodes have an added colour property, and each colour coalesces according to a colour specific, time dependent case. Nodes of non-identical colour can coalesce iff at least one of them is the last remaining node of a given colour. Different colours

correspond to different clades, each behaving under its own growth function.
Similar models have been used in epidemiology to track outbreaks [19], or transmission chains [20]. These models are often referred to as referred to as structured coalescent process, effectively adding a colour property to the vertices of phylogenies.

1) Model

A given genealogy $\mathbf{g} = (V_\mathbf{g}, E_\mathbf{g}, t_\mathbf{g})$ is an incomplete, empirical sample of the underlying process.
It consists of nodes $V_\mathbf{g}$, directed edges $E_\mathbf{g}$, and node labels $t_\mathbf{g}$ corresponding to event times.
The genealogy $\mathbf{g}$ shall be indexed by an index set $S = 1 \leq i \leq N \subset \mathbb{N}$, with $Y \subset S$ corresponding to coalescent events and $I \subset S$ corresponding to sampling events.
For convenience, assume that all edges are in the forwards time direction, i.e.:

$$\forall k, l \in S : (k, l) \in E_\mathbf{g} \Rightarrow t_k < t_l$$

Furthermore, all event times are ordered in descending (backwards) time order, with the first event corresponding the the most recent sample

$$\forall k, l \in S : k < l \Rightarrow t_k > t_l$$

Under the assumption that $\mathbf{g}$ is a genealogy of a given sample, with each edge in $E_\mathbf{g}$ there is an associated unobserved set of individuals descending from one another. At some point along an edge from one lineage to another, the lineage can undergo a colour change, and become the most recent ancestor of a diverging clade. This event corresponds to this lineage somehow gaining advantage over other lineages, be it a bacterium gaining resistance against a drug, or a strain of a virus invading a completely susceptible population.

Definition III.1 (Multiple Lineage Coalescent)**.** Given M colours, M population size functions $\alpha \triangleq \{\alpha_j(t)\}_{1 \leq j \leq M}$. Let $Y(t)$ be a continuous time markov chain (CTMC) with the state space:

$$\Sigma = \{\mathbf{s} \in \mathbb{Z}_+ : |\mathbf{s}| \geq 1\} \tag{18}$$

and the transition rates

$$\mathbf{s} \to \mathbf{s} - \mathbf{e_j} \qquad \binom{s_j}{2} \alpha_j^{-1}(t) \qquad 1 \leq j \leq M \tag{19}$$

$$\mathbf{s} \to \mathbf{s} - \mathbf{e_j} + \mathbf{e_k} \qquad \delta_{1,j} \beta s_k \qquad 1 \leq j, k \leq M \tag{20}$$

Where β is an unknown rate

The interpretation of this model in backwards (coalescent) time is that each node corresponds to a single specific clade (colour). Nodes of the same clade coalesce at i.i.d rates, according to a clade specific growth functions, until reaching the MRCA of given clade. The MRCA then changes type (colour) to that of a different clade. Under the hypothesis, along each of the edges from the parent of a clade MRCA to the MRCA lies a point in that characterises the time of divergence of the clade, after which the clade starts to undergo clonal expansion.

2) Inference

Assume that we have been given a genealogy as described in previous section. The number of individual clades M, the effective population size functions α, and the placement and divergence events in descending order $\Xi = \{\xi_i\}_{1 \leq i \leq M}$, are unknown. In effect we seek to infer the posterior using the following factorisation:

$$\mathcal{L}(\alpha, \Xi, M \mid \mathbf{g}) \propto \mathcal{L}(\mathbf{g} \mid \Xi, M, \alpha) \mathcal{L}(\alpha \mid \Xi, M) \mathcal{L}(\Xi, M) \tag{21}$$

$$\mathcal{L}(\mathbf{g} \mid \Xi, M, \alpha) \tag{22}$$

Definition III.2 (Divergence Event)**.** A divergence event ξ associated with a clade is a tuple:

$$\xi = (e', \tau) \tag{23}$$

where τ denotes the time of the divergence event, and $e' = (v^{pa}, v^{MRCA}) \in E_\mathbf{g}$ denotes the edge along which the divergence event happens, with v^{MRCA} being the MRCA of the newly diverging clade.

In order to identify the likelihood $\mathcal{L}(\mathbf{g} \mid \Xi, M, \alpha)$, the genealogy needs to be first partitioned into separate subtrees, each corresponding to an individual clade.

Definition III.3 (Descendant Set)**.**

$$\mathcal{D}(i) \triangleq \{j \in S \mid j \text{ is a descendant of } i\} \tag{24}$$

As lineages within a clade can only coalesce with one another, the coalescence rate is independent of other clades, and the time of divergence has been given, the total likelihood can be derived as the product of individual clade specific likelihoods.

Definition III.4 (Lineage Subtree)**.** A subtree $W(\xi_i)$ associated with a divergence event ξ_i consists of all vertices, edges and times associated with a given clade.

$$W(\xi_i) \triangleq \mathcal{D}(v_i^{pa}) \setminus \bigcup_{j<i} (\mathcal{D}(v_j^{pa})) \tag{25}$$

A final step before deriving clade specific likelihoods is required, that is, the divergence times have to be included along with the times contained within the clade subtree. These in effect behave identically to sampling events

Denote the set of all divergence times associated with a Lineage Subtree $W(\xi_j)$ $\tau_{W(\xi_j)}$

$$\tau_{W(\xi_j)} \triangleq \tau_i \in \xi_i : v_i^{pa} \in W(\xi_j), \quad \forall 1 \leq i \leq M$$

Finally define the set of all times associated with a particular clade $T(\xi_j) = \tau_{W(\xi_j)} \cup t_{W(\xi_j)}$, indexed by set $S_{T(\xi_j)} = \{1...|T(\xi_j)|\}$ such that

$$\forall t_k, t_l \in T(\xi_j), \quad k > l \Rightarrow t_k \geq t_l$$

Let k_{i,ξ_j} denote the number of extant individuals corresponding to clade subtree ξ_j at during the time interval $[t_{i-1}, t_i]$. The total likelihood is equal to:

$$\mathcal{L}(\mathbf{g} \mid \Xi, M, \alpha) = \prod_{j=1}^{M} \mathcal{L}(T(\xi_j) \mid \alpha_j) \tag{26}$$

$$\mathcal{L}(T(\xi_j) \mid \alpha_j) = \prod_{t_i \in T(\xi_j)} \left(\mathbb{1}_{t_Y}(t_i) \frac{\binom{k_{i,\xi_j}}{2}}{\alpha_j(t_i)} + \mathbb{1}_{t_I \cup \tau_{W(\xi_j)}}(t_i) \right) \times \exp\left(- \int_{t_{i-1}}^{t_i} \frac{\binom{k_{i,j}}{2}}{\alpha_j(\tau)} d\tau \right) \tag{27}$$

3) Choice of Population Size Functions

While the parent clade is assumed to be at equilibrium, the population size functions of the diverging clades have to satisfy several properties.

First we introduce the variable $\tau = -t + T_{max} - T_{div}$ whch denotes time relative to a divergence event of a given clade, where T_{div} denotes the divergence time and T_{max} denotes the time of the most recent sample. In our model we assume that the diverging subpopulation only appears after the divergence event, and as such it is required that at $\tau - 0$ $\alpha(\tau) = 0$. Furthermore, we're a for a monotone decreasing function in τ, that exhibits saturating behaviour as τ grows large. Initially, functions exhibiting a period of exponential growth were investigated, however these were hard to justify and were ill-posed numerically. Hence we arrived at the following function

$$\alpha(\tau) = K \frac{r\tau^2}{1 + r\tau^2} \tag{28}$$

Where K is the carrying capacity and r is the growth rate. This function exhibits saturating behaviour, symmetry around zero, and numerically stable behaviour due to not relying on time-offsetting exponentials and having a more gradual decay around zero.

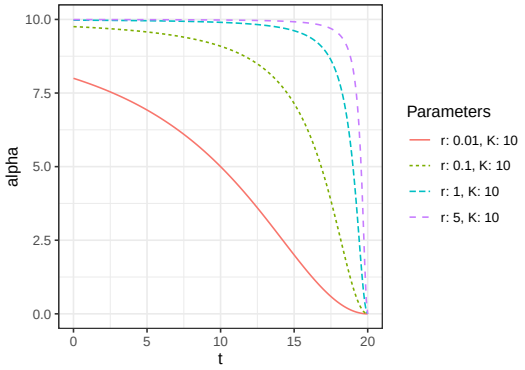


FIGURE 1: $\alpha(\tau)$ under different parameters, with $T_{div} = 20$

The integral of the reciprocal $\alpha^{-1}(\tau)$ under this formulation is then given by

$$\int_t^{t+s} \alpha^{-1}(\tau)\, d\tau = \frac{1}{K} \left[-\frac{1}{r(\tau - T_{max} + T_{div})} + \tau - T_{max} + T_{div} \right]_t^{t+s} \tag{29}$$

As $t + s$ approaches the divergence time relative to the most recent sample $T_{div} - T_{max}$, the rate integral 29 approaches infinity, and as such all coalescence within a clade happens before time of divergence with probability one.

IV. RESULTS

A. IMPLEMENTATION NOTES

The framework described is being implemented as an R package, featuring the ability to simulate genealogies under the model we propose, compute model likelihoods for varying parameters. and to perform inference under same model using MCMC. Please note, as this is still unpublished and work in progress, the repository is currently private access can be granted upon request. Once this has been submitted for publication and completed, the package will be released as open source.

B. EXPONENTIAL GROWTH

1) MCMC Inference

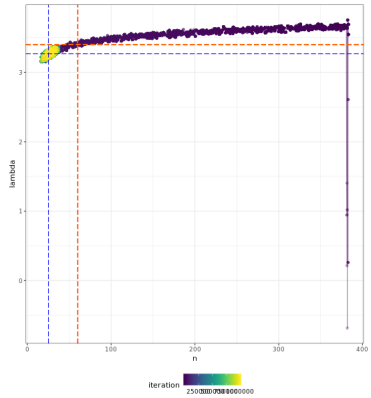
First thing that has been implemented and tested was a basic framework that enables simulating genealogies under the standard inhomogenous coalescent process, and the computation of likelihoods under the standard inhomogenous coalescent model. Along with this, standard metropolis-hastings MCMC routine was implemented

and subsequently tested by simulating a genealogy undergoing exponential growth, and performing inferences on the rate parameter and the maximum population size parameter, effectively replicating results in [15].

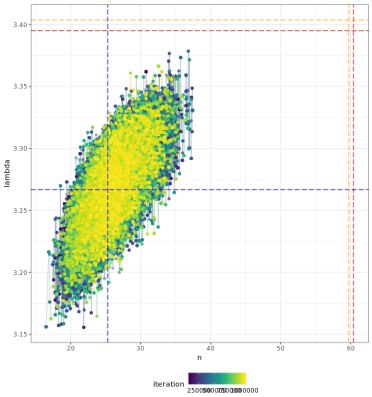
This example used the relative population size function $\alpha(t) = N * \exp(-\lambda t)$, and consisted of 100 sampling events between 0 and 10 years before present. Accordingly, a genealogy has then been simulated with a rate parameter λ and final size parameter N drawn from a uniform densities on intervals $[0.1, 10]$, and $[1, 100]$ respectively.

A Metropolis-Hastings MCMC scheme was then used to infer the parameters λ and N. An exponential distribution rate equal to one was chosen as the prior for λ, as well as for N.

One million iterations were used and the first One hundred thousand discarded as burn in time. To further validate the fit, both a maximum likelihood (MLE) and maximum *a-posteriori* (MAP) estimates were computed and plotted against the posterior marginals inferred by the MCMC.

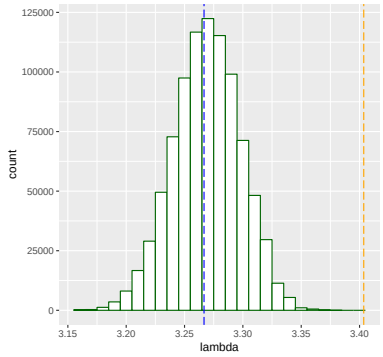


(a)

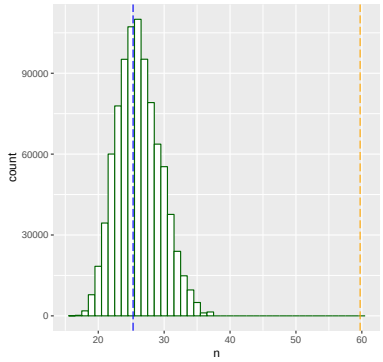


(b)

FIGURE 2: Trace plots for the markov chain. Red lines denote true parameter values. MLE marked by orange lines. MAP marked by blue lines. 2a Shows the entire trace of the chain. 2b Shows the trace with the first 100000 iterations discarded. Colour gradient corresponds to the iteration number, with more recent iterations in yellow.



(a)



(b)

FIGURE 3: Histograms of the posterior marginals. MLE marked by orange line. MAP marked by blue line. 3a λ marginal 3b N marginal

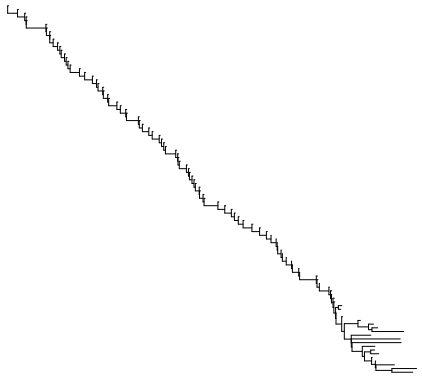


FIGURE 4: The simulated genealogy used for this example.

,

C. COALESCENT WITH LOCAL POPULATION STRUCTURE

1) Phylogeny Simulation

The package in development is currently capable of simulating phylogenies based on the coalescent with local population structure proposed in this work. Phylogenies can be simulated with an arbitrary number of diverging subpopulations and clearly visualised.

The package is also capable of rapidly computing likelihoods for the arbitrary phylogenies and parameters. The likelihood computation is partly implemented in C++ for increased perfromance.

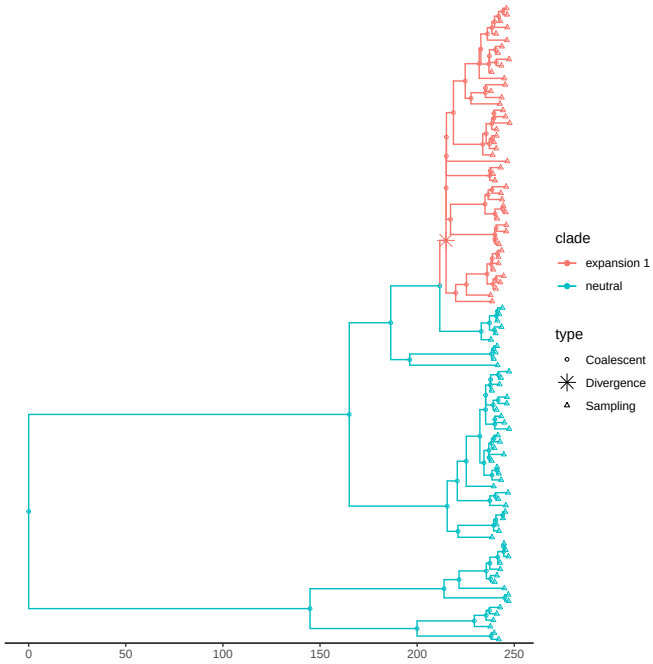


FIGURE 5: A simulated genealogy with one clonal expansion

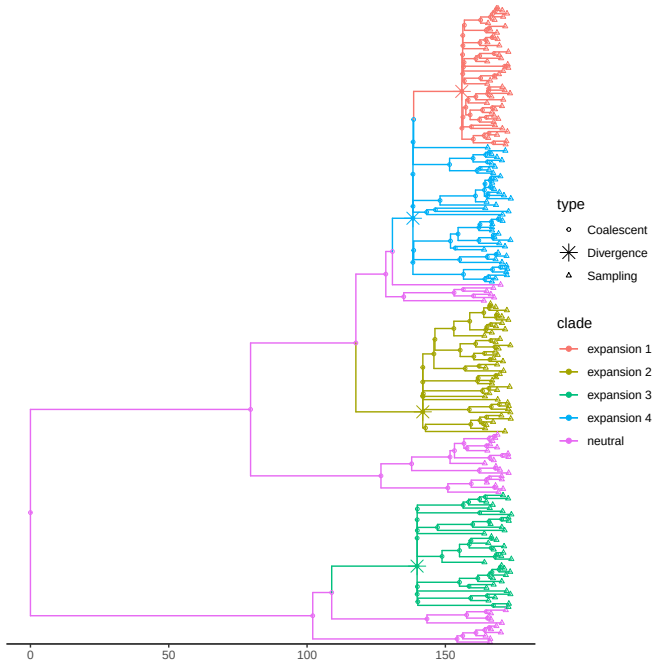


FIGURE 6: A simulated genealogy with four clonal expansions

2) MCMC inference

At the moment we are in the process of implementing, and valdiating the MCMC routine for inference under coalescent with local population structure, and with a fixed number of divergence events. Currently there is a bug in the proposal move that changes the location of the divergence event, specifically, when it's supposed to traverse the root of the tree. As such we are unable to present reliable results for

V. DISCUSSION & FUTURE WORK
VI. DISCUSSION

In this work we proposed a framework to enable local phylodynamic inference, with the aim to facilitate usage of genomic data for outbreak detection. We implemented this model in an R package which we plan to release as open source on github once completed. To our knowledge this is a novel model, and there are no other models designed to capture the process of clonal expansions in genealogies.

At the moment we are still experiencing implementation issues with the MCMC routine for the coalescent with local population structure, and these need to be addressed before thoroughly testing the MCMC inference on simulated data.

The effect of priors and of the choice of the population size function family for divergence events is not understood and has not been properly investigated. It would be desirable to investigate what effect might different function families and parameter priors have on inference under our model.

It would be worth investigating the analytic properties of model proposed, and to see whether it can be shown to be consistent with existing models in population genetics in the limit, under appropriate assumptions.

VII. FUTURE WORK

Once work on the MCMC routine for a fixed number of divergence events is completed and the routine is validated, we will move onwards to expand this to an arbitrary number of divergence events.

Having a variable amount of divergence events leads to the dimensionality of the problem not being fixed, and therefore will require the use of Reversible Jump MCMC (RjMCMC) [21, 22], which computationally intensive.

If the performance of this framework in this setting is satisfactory, a further comparison with existing methods such as [7, 10] would be interesting and desirable.

Finally, after the validation of the method is complete, we plan on using this model to analyse real world datasets, ideally those provided by Public Health England to see if outbreaks can be detected.

We plan on releasing the R package implementing our methodology as open source.

References

References

[1] J. M. Smith et al. "How clonal are bacteria?" In: *Proceedings of the National Academy of Sciences* 90.10 (May 15, 1993), pp. 4384–4388. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.90.10.4384. URL: http://www.pnas.org/cgi/doi/10.1073/pnas.90.10.4384 (visited on 07/29/2020).

[2] Brian G. Spratt et al. "Displaying the relatedness among isolates of bacterial species – the eBURST approach". In: *FEMS Microbiology Letters* 241.2 (2004). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1016/j.fem pp. 129–134. ISSN: 1574-6968. DOI: 10.1016/j.femsle.2004.11.015. URL: https://onlinelibrary.wiley.com/doi/abs/10.1016/j.femsle.2004.11.015 (visited on 07/29/2020).

[3] Christophe Fraser, William P. Hanage, and Brian G. Spratt. "Neutral microepidemic evolution of bacterial pathogens". In: *Proceedings of the National Academy of Sciences of the United States of America* 102.6 (Feb. 8, 2005), pp. 1968–1973. ISSN: 0027-8424. DOI: 10.1073/pnas.0406993102. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC548543/ (visited on 07/29/2020).

[4] Alice Ledda et al. "Re-emergence of methicillin susceptibility in a resistant lineage of Staphylococcus aureus". In: *Journal of An-*

timicrobial Chemotherapy 72.5 (May 1, 2017). Publisher: Oxford Academic, pp. 1285–1288. ISSN: 0305-7453. DOI: 10.1093/jac/dkw570. URL: https://academic.oup.com/jac/article/72/5/1285/2930201 (visited on 07/29/2020).

[5] Matthew T. G. Holden et al. "A genomic portrait of the emergence, evolution, and global spread of a methicillin-resistant Staphylococcus aureus pandemic". In: *Genome Research* 23.4 (Apr. 2013), pp. 653–664. ISSN: 1549-5469. DOI: 10.1101/gr.147710.112.

[6] Li-Yang Hsu et al. "Evolutionary dynamics of methicillin-resistant Staphylococcus aureus within a healthcare system". In: *Genome Biology* 16.1 (Apr. 23, 2015), p. 81. ISSN: 1465-6906. DOI: 10.1186/s13059-015-0643-z. URL: https://doi.org/10.1186/s13059-015-0643-z (visited on 07/29/2020).

[7] Erik M. Volz et al. "Identification of Hidden Population Structure in Time-Scaled Phylogenies". In: *Systematic Biology* (). DOI: 10.1093/sysbio/syaa009. URL: https://academic.oup.com/sysbio/advance-article/doi/10.1093/sysbio/syaa009/5734655 (visited on 07/01/2020).

[8] Bethany L. Dearlove and Simon D. W. Frost. "Measuring Asymmetry in Time-Stamped Phylogenies". In: *PLOS Computational Biology* 11.7 (July 6, 2015). Publisher: Public Library of Science, e1004312. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1004312. URL: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004312 (visited on 07/29/2020).

[9] Vegard Eldholm et al. "Four decades of transmission of a multidrug-resistant Mycobacterium tuberculosis outbreak strain". In: *Nature Communications* 6.1 (May 11, 2015). Number: 1 Publisher: Nature Publishing Group, p. 7119. ISSN: 2041-1723. DOI: 10.1038/ncomms8119. URL: https://www.nature.com/articles/ncomms8119 (visited on 07/29/2020).

[10] Joëlle Barido-Sottani, Timothy G. Vaughan, and Tanja Stadler. "A Multitype Birth–Death Model for Bayesian Inference of Lineage-Specific Birth and Death Rates". In: *Systematic Biology* 69.5 (Sept. 1, 2020). Publisher: Oxford Academic, pp. 973–986. ISSN: 1063-5157. DOI: 10.1093/sysbio/syaa016. URL: https://academic.oup.com/sysbio/article/69/5/973/5762626 (visited on 08/28/2020).

[11] J. F. C. Kingman. "The coalescent". In: *Stochastic Processes and their Applications* 13.3 (Sept. 1, 1982), pp. 235–248. ISSN: 0304-4149. DOI: 10.1016/0304-4149(82)90011-4. URL: http://www.sciencedirect.com/science/article/pii/0304414982900114 (visited on 07/30/2020).

[12] R. C. Griffiths et al. "Sampling theory for neutral alleles in a varying environment". In: *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* 344.1310 (June 29, 1994). Publisher: Royal Society, pp. 403–410. DOI: 10.1098/rstb.1994.0079. URL: https://royalsocietypublishing.org/doi/10.1098/rstb.1994.0079 (visited on 08/28/2020).

[13] Jotun. Hein, Mikkel H. Schierup, and Carsten. Wiuf. *Gene Genealogies, Variation and Evolution: A primer in coalescent theory:* Oup Oxford, Dec. 9, 2004. ISBN: 978-0-19-154615-0. URL: https://www.dawsonera.com:443/abstract/9780191546150.

[14] O G Pybus, A Rambaut, and P H Harvey. "An integrated framework for the inference of viral population history from reconstructed genealogies." In: *Genetics* 155.3 (July 2000), pp. 1429–1437. ISSN: 0016-6731. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1461136/ (visited on 08/28/2020).

[15] Alexei J. Drummond et al. "Estimating Mutation Parameters, Population History and Genealogy Simultaneously From Temporally Spaced Sequence Data". In: *Genetics* 161.3 (July 1, 2002). Publisher: Genetics Section: INVESTIGATIONS, pp. 1307–1320. ISSN: 0016-6731, 1943-2631. URL: https://www.genetics.org/content/161/3/1307 (visited on 07/02/2020).

[16] Mandev S. Gill et al. "Improving Bayesian Population Dynamics Inference: A Coalescent-Based Model for Multiple Loci". In: *Molecular Biology and Evolution* 30.3 (Mar. 1, 2013). Publisher: Oxford Academic, pp. 713–724. ISSN: 0737-4038. DOI: 10.1093/molbev/mss265. URL: https://academic.oup.com/mbe/article/30/3/713/1041171 (visited on 06/30/2020).

[17] Mandev S. Gill et al. "Understanding Past Population Dynamics: Bayesian Coalescent-Based Modeling with Covariates". In: *Systematic Biology* 65.6 (Nov. 1, 2016). Publisher: Oxford Academic, pp. 1041–1056. ISSN: 1063-5157. DOI: 10.1093/sysbio/syw050. URL: https://academic.oup.com/sysbio/article/65/6/1041/2281638 (visited on 06/30/2020).

[18] Radek Erban, S Jonathan Chapman, and Philip K Maini. "A PRACTICAL GUIDE TO STOCHASTIC SIMULATIONS OF REACTION-DIFFUSION PROCESSES". In: *DIFFUSION PROCESSES* (), p. 35.

[19] Nicola De Maio, Chieh-Hsi Wu, and Daniel J. Wilson. "SCOTTI: Efficient Reconstruction of Transmission within Outbreaks with the Structured Coalescent". In: *PLOS Computational Biology* 12.9 (Sept. 28, 2016). Publisher: Public Library of Science, e1005130. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1005130. URL: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1005130 (visited on 07/22/2020).

[20] Xavier Didelot et al. "Genomic Infectious Disease Epidemiology in Partially Sampled and Ongoing Outbreaks". In: *Molecular Biology and Evolution* 34.4 (Apr. 1, 2017). Publisher: Oxford Academic, pp. 997–1007. ISSN: 0737-4038. DOI: 10.1093/molbev/msw275. URL: https://academic.oup.com/mbe/article/34/4/997/2919386 (visited on 07/15/2020).

[21] Peter J. Green. "Reversible jump Markov chain Monte Carlo computation and Bayesian model determination". In: *Biometrika* 82.4 (Dec. 1, 1995). Publisher: Oxford Academic, pp. 711–732. ISSN: 0006-3444. DOI: 10.1093/biomet/82.4.711. URL: https://academic.oup.com/biomet/article/82/4/711/252058 (visited on 07/14/2020).

[22] Y. Fan and S. A. Sisson. "Reversible jump Markov chain Monte Carlo". In: *arXiv:1001.2055 [stat]* (Jan. 12, 2010). arXiv: 1001.2055. URL: http://arxiv.org/abs/1001.2055 (visited on 07/13/2020).