A scalable analytical approach from bacterial genomes to epidemiology

Xavier Didelot[1] and Julian Parkhill[2]

1 School of Life Sciences and Department of Statistics, University of Warwick, Coventry, UK
2 Department of Veterinary Medicine, University of Cambridge, Cambridge, UK

Keywords: bacterial genomics, infectious disease epidemiology, recombination, dated phylogeny

Summary

Recent years have seen a remarkable increase in the practicality of sequencing whole genomes from large numbers of bacterial isolates. The availability of this data has huge potential to deliver new insights into the evolution and epidemiology of bacterial pathogens, but the scalability of the analytical methodology has been lagging behind that of the sequencing technology. Here we present a step-by-step approach for such large-scale genomic epidemiology analyses, from bacterial genomes to epidemiological interpretations. A central component of this approach is the dated phylogeny, which is a phylogenetic tree with branch lengths measured in units of time. The construction of dated phylogenies from bacterial genomic data needs to account for the disruptive effect of recombination on phylogenetic relationships, and we describe how this can be achieved. Dated phylogenies can then be used to perform fine-scale or large-scale epidemiological analyses, depending on the proportion of cases for which genomes are available. A key feature of this approach is computational scalability, and in particular the ability to process hundreds or thousands of genomes within a matter of hours. This is a clear advantage of the step-by-step approach described here. We discuss other advantages and disadvantages of the approach, as well as potential improvements and avenues for future research.

1. Introduction

Over the past decade, the cost and time required to sequence whole bacterial genomes has reduced dramatically [1]. Sequencing is frequently applied to many or all isolates in local outbreaks, or to a high proportion of cases in more endemic situations, as well as large retrospective and longitudinal collections. This genomic data has huge potential to deliver new insights into the evolution and epidemiology of bacterial pathogens, which can lead to better control measures. However, the lack of scalable methodology for analysis of this genomic data represents an important bottleneck for the realisation of their full potential.

A gold standard for the analysis of pathogen genomic data has been set by the integrated phylogenetic frameworks implemented for example in BEAST [2] and BEAST2 [3]. These phylodynamic tools were originally conceived for viral genetics and are still mostly used for that purpose, but have also been increasingly applied to bacterial genomic data [4]. One of the strengths of these tools is that they can infer a dated phylogeny by combining the genomic data with the dates of isolation, resulting in estimates for the dates of the common ancestors in the phylogeny. Such dated phylogenies are extremely useful to draw epidemiological interpretations from the genomic data, as we will see. Another advantage of the integrated phylogenetic frameworks is that they include a number of powerful extensions, for example to use relaxed clock models [5], to estimate past population dynamics [6], geographical spread [7–9] or transmission between hosts [10,11]. This integrated approach has many natural advantages but also limitations especially in terms of scalability to analyse larger datasets.

These limitations of the integrated approach are especially important in bacterial genomics, where the genomes are orders of magnitude longer than in viral genetics and often subject to recombination. The ClonalOrigin model [12] of bacterial evolution has been integrated into BEAST2 [13], but the resulting algorithm is too computationally intense to be applied to whole genome datasets. Here we present an alternative step-by-step approach.

*Author for correspondence (xavier.didelot@warwick.ac.uk).

The step-by-step approach is illustrated in Figure 1. In the first step, a phylogeny is constructed from a genomic alignment in a way that accounts for recombination events. In the second step, this phylogeny is dated. In the third step, the dated phylogeny is interpreted in terms of a number of epidemiological properties. Many software packages are available to perform each of these steps, including but not limited to the ones named in Figure 1, although it is worth noting that many of these tools have emerged only in the past few years, and so are still work in progress and expected to improve in the near future. In this article we review each of the steps of this approach in turn. We also pay special attention to the 'cracks' between the steps, since these are often ignored in articles that focus on each of the steps rather than the whole step-by-step approach. Finally, we demonstrate the usability of this approach by applying it to a complete collection of *Staphylococcus aureus* ST239 genomes.

2. Recombination-aware phylogenetic analysis

Even a relatively low amount of recombination can invalidate the results of phylogenetic tools if not accounted for [14,15]. It is therefore essential to detect recombination events to correctly reconstruct the clonal genealogy, that is the phylogenetic relationship between genomes when the ancestral lines of recipient cells rather than donor cells are followed for each ancestral recombination events. Special phylogenetic methods have been developed for this purpose, including Gubbins [16] and ClonalFrameML [17] which is based on the ClonalFrame model [18]. However, these tools are often underexploited, typically to build a recombination-corrected tree without paying attention to the recombination events and regions that have been detected.

A lot can be learnt from studying the inferred recombination events themselves. Recombination is useful to help us understand how species are being formed [19] and the population structure within species, especially when the origin of recombination events is being investigated [20]. These recombination patterns often reflect important driving evolutionary forces such as ecology [21], adaptation [22] or selective pressures [23]. For example in *Streptococcus pneumoniae*, recombination events have been shown to be driven by antibiotic usage in a localised dataset [24] and by immune pressure in a global collection of the PMEN1 lineage [25]. The latter study also represents a good example of how the temporal signal can become much clearer once recombination is correctly accounted for [25,26]. Recombination is also useful for the analysis of genome-wide associations between genotypes and phenotypes, since it separates new genetic variants from their original genomic background [27].

Accounting for recombination when reconstructing phylogenies is an important starting point for many epidemiological studies. A method often used is to extract from the genomic alignment the sites that have not been affected by recombination and to build a phylogeny using these sites only. Both Gubbins and ClonalFrameML are often used in this way, to create a recombination-free alignment which is then passed on to BEAST. However, this method works only if relatively few recombination events happened throughout the tree. For example, consider the simulated dataset shown in Figure 2. The true clonal genealogy is shown in Figure 2A and the true recombination events that happened on each of the branches are shown in Figure 2B. These data were simulated using a standard coalescent model for the phylogeny [28], a strict clock model of mutation with rate $\theta/2=0.005$ per site, a model of recombination coming from external sources [18] with initiation rate $\varrho/2=0.001$ per site, average length of recombination $\delta=1500$bp and distance of the source $\nu=0.05$. For clarity we used a relatively small dataset of 20 sequences of 100,000bp each. In this simulated dataset, there was not a single site that was not affected by recombination on at least one of the branches. On the other hand, every branch had some sites unaffected by recombination (Figure 2B).

We applied ClonalFrameML [17] to this dataset using a PhyML tree [29] as starting point. The reconstructed clonal genealogy is shown in Figure 2C and the inferred recombination events are shown in Figure 2D, and they are in very good agreement with the true simulated tree and events shown in Figures 2A and 2B. ClonalFrameML correctly inferred that there was not a single site unaffected by recombination on at least one of the branches. Therefore an alignment containing only the non-recombinant sites would contain no sites, and could not be used as a starting point for further analysis. On the other hand, the inferred clonal genealogy shown in Figure 2C can be used in our proposed step-by-step approach. It has the same topology as the true clonal genealogy (Figure 2A) and very similar branch lengths, with a weighted Robinson-Foulds distance [30] of 0.005 between the true and ClonalFrameML trees. Gubbins [16] was also applied to this dataset using RAxML [31] as a tree builder. The correct topology was inferred, with a weighted Robinson-Foulds distance of 0.03 between the true and Gubbins trees.

3. Dating the ancestors in a phylogeny

Once a recombination-corrected tree has been reconstructed, it is possible to study the temporal signal in this tree and to date the common ancestors in the tree. Multiple software tools have recently been developed to perform dating on a phylogeny, including BactDating [26] which is specifically aimed at bacterial genomes, but also LSD [32], treedater [33] and TreeTime [34]. BactDating uses Bayesian statistics, whereas treedater and TreeTime are based on maximum likelihood, which is identical to a Bayesian maximum a-posteriori (MAP) approach assuming a uniform prior on dates as previously proposed [35]. It is often important to use a relaxed clock model in this step that allows the evolutionary rate to vary between lineages [5]. An additive relaxed

clock model has recently been developed which is more biologically realistic and leads to better dating of pathogen phylogenies than previous relaxed clock model [36].

In our proposed step-by-step approach, the reconstruction of a dated phylogeny and its epidemiological interpretation are separated. One disadvantage of this is that the prior (or lack-of) on dates used to reconstruct the dated phylogeny is not the same as the one that would be implied by the epidemiological models used in subsequent analyses. This statistical issue could be resolved for example by considering the difference in tree distribution between the models used for the dating and the epidemiology and applying an importance sampler to correct for this difference [37]. However, this difference is often small enough to be ignored in practice, especially if the method used to build the dated phylogeny was based on the likelihood only, or if a mild prior was used such as the coalescent with constant population size [28].

To illustrate this, we simulated five years of an outbreak model [38] with within-host diversity $N_e g$=0.25 year, basic reproduction number R_0=2, generation time distribution Exponential(1) in years and sampling proportion π=0.1. A total of 59 cases were sampled in this outbreak, with the samples being related as shown in Figure 3A. We applied a strict clock model to this dated phylogeny with a rate μ=5 substitutions per year which is of the same order of magnitude as many bacterial pathogens [39]. This undated phylogeny was then used, along with the known dates of sampling, to infer a dated phylogeny using BactDating [26] with prior set to the coalescent with constant population size [28]. This prior is very different from the outbreak model that was used to generate the phylogeny [38], which is not coalescent due to the host structure and where the population size is clearly growing since the reproduction number was greater than one. Figure 3B shows the inferred dating for this tree, which is in good agreement with the correct dates from the top part despite the complete difference between the epidemic model used for simulation and the coalescent model used for inference.

At the same time as dating is performed, the substitution rate is typically estimated which provides a useful value to compare with previous estimates [39] in order to make sure that the dating is working as expected. Statistical methods can also be used to ensure that the temporal signal is significant, for example by comparing the fit of the data when the correct sampling dates are used against when all the dates are forced equal [40], or using a permutation test on the sample dates [41]. These methods require to perform several runs of the dating method, and it is therefore useful for this to be as fast as possible, which is achieved in our step-by-step method by separating the phylogenetic inference from the dating.

Furthermore, the root of the phylogeny is typically estimated during the dating step, since the trees generated by standard phylogenetic tools are not rooted whereas dated trees are always rooted by definition, with the date of the root being the date of the last common ancestor of the whole sample. If the root has already been determined robustly, for example using one or ideally several closely related outgroups [42], then this information can be preserved during the dating. If on the other hand the root is undetermined, or arbitrarily selected for example using the midpoint method [43], then the fact that dating the phylogeny simultaneously performs rooting provides an additional reason for dating the tree, which becomes much more informative in terms of epidemiology once it is dated and rooted.

4. From dated phylogeny to epidemiology

A dated phylogeny is very useful to learn about the epidemiology of the bacteria under study, and sometimes the dating directly provides answers to questions of interest beyond the age of pathogens [44]. For example, several antibiotic resistant lineages have been dated to have emerged around the time when the corresponding antibiotics were started to be used, highlighting the link between consumption and resistance [45,46]. As another example, the dating of the common ancestors between pairs of *Clostridium difficile* patients in a hospital allowed to rule out transmission for many pairs and to conclude that nosocomial transmission was less frequent than previous thought [47].

It can often be useful to identify clusters of significantly similar genomes in a dataset. The most commonly used approach is to use a separate dedicated algorithm that uses the genomic data for this purpose, such as HierBAPS [48], fastbaps [49] or PopPunk [50], and overlay the results of this clustering analysis onto the phylogeny using colours for example. Another approach is to use additional non-genomic data to do the clustering given the phylogeny, as performed for example by AdaptML [51], treebreaker [52] and treeSeg [53]. Finally a third option is to try and identify directly on the dated phylogeny the lineages that seem to be ruled by different dynamics, for example using treestructure which does not rely on an explicit phylodynamic model [54] or CaveDive [55] which is focused on the detection of clonal expansions.

The dated phylogeny can also be used as a starting point for further analysis. In particular, past variations in the bacterial population size have a direct effect on the shape of the dated phylogeny, so that the population size through time can be estimated and presented as a skyline plot [56]. The methodology for performing such an analysis was originally developed within BEAST which simultaneously estimates the dated phylogeny [6], but for our step-by-step approach we need to estimate the demographic function from a dated phylogeny, and several software tools have recently been released for this purpose including phylodyn [57,58], skygrowth [59] and mlesky [60]. Beyond a simple model of varying population size, it is also possible to fit an epidemiological

compartmental model such as the susceptible-infected-recovered model [61], and therefore to estimate the parameters of this model such as the transmission rate or removal rate. Such an inference can be achieved by formulating a structured coalescent model that corresponds to the compartmental model [62,63]. Existing software for fitting such a model to a given dated phylogeny include rcolgem [64] and phydyn [65]. The same methods based on the structured coalescent can also be applied to a dated phylogeny in order to reconstruct past geographical migrations [9], although such phylogeographic inference is much more often based on discrete trait analysis, for example using the ace command from the R package ape [66] or in the NextStrain platform [67]. The worldwide spread of the current pandemic of *Vibrio cholerae* has been described using such techniques [68,69].

When the genomes are densely sampled within an epidemic, it can be useful to try and reconstruct the transmission tree of who infected whom [70]. Within-host diversity and evolution is significant for many bacterial pathogens which blurs the relationships between transmission tree and phylogeny [71]. However, TransPhylo can infer the transmission tree from a dated phylogeny in a way that accounts for within-host evolution [38,72,73]. Significant uncertainty typically remains in the inferred transmission tree, which is captured by the use of Bayesian statistics within TransPhylo. More precise inference can sometimes be obtained by combining the genomic inference with epidemiological data [74].

A drawback of separating the dating step from the interpretation step is that the uncertainty in dating is typically not passed on to the epidemiological analysis. This can be achieved by running on multiple samples from the posterior of dated phylogeny and averaging the results [75], or reweighting according to the posterior probability in the epidemiological analysis [37], but in practice the phylogenetic uncertainty is usually not accounted for. However, this is not often a significant issue in practice. To illustrate this, we simulated a dataset for a small outbreak with just ten cases, using an epidemic model [38] with basic reproduction number R_0=1, within-host diversity $N_e g$=0.25 year, mean generation time of 1 year, sampling proportion of π=0.5 and a strict clock model with rate μ=5 substitutions per year. The dated phylogeny was inferred using BactDating [26] and we extracted the first (after burnin) and the last trees sampled by the MCMC, as shown in Figures 4A and 4C. We then reconstructed the transmission events using TransPhylo [38] separately for each of these two dated trees, as shown in Figures 4B and 4D. In spite of small differences in the two dated phylogenies, the inferred results in terms of transmission chains were very similar.

5. Example of application

To illustrate the use of the step-by-step approach from bacterial genomes to epidemiology, we apply it to a state-of-the-art dataset, using only a standard laptop computer and paying particular attention to the time taken by each step. We collected all available genomes of *Staphylococcus aureus* ST239 (Table S1). This collection is made of 521 assembled genomes, only small subsets of which had been comparatively analysed in previous studies [76–79]. The genomes were collected between 1982 and 2010 from all parts of the world (451 from Asia, 46 from Europe, 18 from Americas, 2 from Africa, 2 from Oceania and 2 unknown). All genomes were aligned using MuMMER v3.1 [80] against the reference genome TW20 which is a member of ST239 [81] and therefore included in the collection. This resulted in a reference-anchored alignment that took only a few minutes to generate, since each pairwise alignment against the reference genome can be performed in parallel. Alternatively, assembly pipelines are often based on reference-based mapping of the sequencing reads, for example using BWA [82] and SamTools [83]. This can also be performed in parallel and results in a similar reference-anchored alignment.

A first phylogeny was built using PhyML v3.3 [29] which took approximately 3 hours. This was used as the starting point to build a recombination-corrected phylogeny using ClonalFrameML v1.12 [17], which took approximately two days to run. The same analysis using Gubbins v2.4.1 [16] gave very similar results, and took approximately one day to run. This step currently represents a clear bottleneck in the application of the step-by-step approach, which should be addressed in the near future through the development of new parallelised algorithms. Significant recombination was found, with a total of 198 recombination events detected throughout the phylogeny. The relative rate of recombination versus mutation was estimated to be R/θ=0.144, meaning that on average mutation events were about 7 times more frequent than recombination events. The mean length of recombination events was estimated to be δ=619 bp which is in good agreement with previous estimates for S. aureus [17,84,85]. The mean distance between donor and recipient was estimated to be ν=0.31%, which corresponds approximately to the distance between ST239 and some of its closest relatives such as CC8 [86]. The relative effect of recombination versus mutation was therefore estimated to be $r/m = R/\theta \times \nu \times \delta = 0.28$, so that 3 to 4 times more substitutions are caused by mutation than by recombination. These results confirm that recombination plays a role in *S. aureus* evolution, although not as dramatic as in some other bacterial pathogens [14,87,88].

We detected a strong temporal signal in the recombination-corrected phylogeny on the basis of a regression analysis of root-to-tip distances against isolation dates (R^2=0.57, $p<10^{-4}$). We therefore computed a dated phylogeny using BactDating v1.1 [26] under the additive relaxed clock model [36]. This step took approximately 3 hours to run for 10^6 MCMC iterations, and the inferred dated phylogeny is shown in Figure 5A. The isolation dates were unknown for 36 of the 521 genomes (Table S1), but BactDating can accommodate this. The evolutionary rate was estimated to be 7.05 substitutions per year throughout the genome, with

credible interval between 6.43 and 7.67. This estimate is in good agreement with several previous estimates in ST239 [76,78] and other lineages of *S. aureus* [46]. The root of the ST239 was estimated to have existed in 1958, with credible interval ranging between 1951 and 1965. This is again in good agreement with previous estimates and coincides with penicillins being increasingly used to treat bacterial infections [76,78,89].

We used the dated phylogeny as input into treestructure v0.1.2 [54] to determine whether there were significant differences in the phylodynamic properties of sublineages within the tree. This analysis took less than a minute to perform, and found no significant differences, which means that the whole tree can be treated as a whole in phylodynamic reconstructions [54]. We therefore applied skygrowth v0.3.1 [59] to the whole dated tree using the maximum a-posteriori method. This analysis took less than a minute and the estimated demographic function is shown in Figure 5B, with an approximately exponential rise of the effective population size between 1960 and 1995, and a plateau between 1995 and 2010. This is in good agreement with previous skyline analyses of ST239 [78,89]. We do not seek to say more about the epidemiological dynamics of ST239 since our aim with this application was to test the applicability of the step-by-step method to a relatively large dataset, rather than study it in detail.

6. Discussion

The step-by-step approach has several drawbacks compared to an integrated approach. A practical disadvantage is that multiple tools need to be applied one after the other, with the need to make sure that the output of one tool is a suitable input for the next tool. The software tools have been developed separately, and format conversion is sometimes required when combining them, which introduces a risk of error being made. Method developers should make every effort to minimize this risk, for example by providing practical examples of source code combining new tools with pre-existing ones, and including verifications in each tool that the input is formatted as expected.

Another concern with the step-by-step approach relates to statistical soundness. In an integrated approach, a complex model is formed by combining multiple simpler models into a consistent whole, for example a model describing how the pathogen population size varied over time, another model describing how these fluctuations affect the genealogy and yet another model describing how mutation and recombination events affect the genomes given the genealogy. Inference is then performed on the combined model, with all uncertainties being accounted for simultaneously and in all directions: for example the uncertainty on a mutation event will feed into the uncertainty on the past population size, and vice-versa. By contrast, in the step-by-step approach, each of the tool makes separate modelling assumptions, which may not always by consistent with each other. An example of this was discussed in Section 3, where the prior used for the reconstruction of a dated phylogeny was not the correct one, but Figure 3 showed that the result can still be correct. Furthermore, the uncertainty can only be passed from one tool to the next in the order that they are being applied, and even in this direction it is frequent to use the best estimate from one method as the starting point of the next, without passing any uncertainty. Again this is not necessarily a problem in practice, as illustrated in Figure 4 where the uncertainty on the phylogeny had little effect on the uncertainty of the transmission tree. From a statistical point of view, the integrated approach therefore represents a gold standard, although statisticians have recently noted that joint inference under a combined model carries the risk that misspecification in any of the model parts can affect estimates from the others in unpredictable ways [90]. Further research is needed on this in the context of genomic epidemiology, as well as research on how to avoid the statistical issues described above with the step-by-step approach.

A key advantage to the step-by-step approach we described is that by breaking down the problem into simple steps, it becomes easier to solve, a strategy often called “divide and conquer” in the computer science literature. The running time is greatly improved compared to an integrated approach, which quickly becomes intractable as more model components are combined into a large model. An example of this concerns the difficulty to integrate recombination into a phylodynamic framework [13]. A similar situation occurs when aligning sequences and building a phylogeny: in principle alignment and phylogeny would benefit from being performed simultaneously [91,92] but in practice this is too computationally challenging. The lower running time of the step-by-step approach also means that it is more scalable to the large numbers of bacterial genomes currently available, and this scalability is probably the main reason for a recent increase in popularity [67,93].

Perhaps even more importantly, a counterintuitive advantage of the step-by-step approach is that it is less automatic than the integrated approach. Although this may seem like a disadvantage, the fact that several software tools have to be applied one after the other brings great benefits. It allows the user to check after each step that the result makes sense before carrying the next step. For example, if a phylogeny is clearly wrong due to contamination during sequencing, there is no point trying to apply dating of the nodes or interpreting the phylogeny in terms of epidemiology. Since each tool is focused on a simpler task, it is easier for the user to check the validity of the assumptions made, and if needed to compare models or the results of several software tools, or apply more complex models since each step is relatively quick. These checks and refinements provide the user with a better understanding of their data and the analysis process, rather than relying on “black-box” or “turn-key” analysis. This is one of the most important advantages of the step-by-step approach, since it creates good conditions for a balanced interpretation of the data and results.

Acknowledgments

We acknowledge funding from the National Institute for Health Research (NIHR) Health Protection Research Unit in Genomics and Enabling Data.

References

1. Loman NJ, Pallen MJ. 2015 Twenty years of bacterial genome sequencing. *Nat. Rev. Microbiol.* **13**, 787–94. (doi:10.1038/nrmicro3565)
2. Suchard MA, Lemey P, Baele G, Ayres DL, Drummond AJ, Rambaut A. 2018 Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evol.* **4**, vey016. (doi:10.1093/ve/vey016)
3. Bouckaert R *et al.* 2019 BEAST 2.5 : An Advanced Software Platform for Bayesian Evolutionary Analysis. *PLoS Comput. Biol.* **15**, e1006650.
4. Ingle DJ, Howden BP, Duchene S. 2021 Development of Phylodynamic Methods for Bacterial Pathogens. *Trends Microbiol.* , 1–10. (doi:10.1016/j.tim.2021.02.008)
5. Drummond AJ, Ho SYW, Phillips MJ, Rambaut A. 2006 Relaxed phylogenetics and dating with confidence. *PLoS Biol.* **4**, e88. (doi:10.1371/journal.pbio.0040088)
6. Drummond AJ, Rambaut A, Shapiro B, Pybus OG. 2005 Bayesian coalescent inference of past population dynamics from molecular sequences. *Mol. Biol. Evol.* **22**, 1185–92. (doi:10.1093/molbev/msi103)
7. Lemey P, Rambaut A, Drummond AJ, Suchard M. 2009 Bayesian phylogeography finds its roots. *PLoS Comput. Biol.* **5**, e1000520. (doi:10.1371/journal.pcbi.1000520)
8. Lemey P, Rambaut A, Welch JJ, Suchard MA. 2010 Phylogeography takes a relaxed random walk in continuous space and time. *Mol. Biol. Evol.* **27**, 1877–1885. (doi:10.1093/molbev/msq067)
9. De Maio N, Wu C-H, O'Reilly KM, Wilson D. 2015 New Routes to Phylogeography: A Bayesian Structured Coalescent Approximation. *PLoS Genet.* **11**, e1005421. (doi:10.1371/journal.pgen.1005421)
10. Hall M, Woolhouse M, Rambaut A. 2015 Epidemic Reconstruction in a Phylogenetics Framework: Transmission Trees as Partitions of the Node Set. *PLOS Comput. Biol.* **11**, e1004613. (doi:10.1371/journal.pcbi.1004613)
11. De Maio N, Wu C-H, Wilson DJ. 2016 SCOTTI: Efficient Reconstruction of Transmission within Outbreaks with the Structured Coalescent. *PLoS Comput. Biol.* **12**, e1005130. (doi:10.1371/journal.pcbi.1005130)
12. Didelot X, Lawson DJ, Darling AE, Falush D. 2010 Inference of homologous recombination in bacteria using whole-genome sequences. *Genetics* **186**, 1435–49. (doi:10.1534/genetics.110.120121)
13. Vaughan TG, Welch D, Drummond AJ, Biggs PJ, George T, French NP. 2017 Inferring ancestral recombination graphs from bacterial genomic data. *Genetics* **205**, 857–870. (doi:10.1534/genetics.116.193425)
14. Didelot X, Maiden MCJ. 2010 Impact of recombination on bacterial evolution. *Trends Microbiol.* **18**, 315–322. (doi:10.1016/j.tim.2010.04.002)
15. Hedge J, Wilson DJ. 2014 Bacterial Phylogenetic Reconstruction from Whole Genomes Is Robust to Recombination but Demographic Inference Is Not. *MBio* **5**, e02158-14. (doi:10.1128/mBio.02158-14.Editor)
16. Croucher NJ, Page AJ, Connor TR, Delaney AJ, Keane JA, Bentley SD, Parkhill J, Harris SR. 2015 Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Res.* **43**, e15. (doi:10.1093/nar/gku1196)
17. Didelot X, Wilson DJ. 2015 ClonalFrameML: Efficient Inference of Recombination in Whole Bacterial Genomes. *PLoS Comput. Biol.* **11**, e1004041. (doi:10.1371/journal.pcbi.1004041)
18. Didelot X, Falush D. 2007 Inference of bacterial microevolution using multilocus sequence data. *Genetics* **175**, 1251–66. (doi:10.1534/genetics.106.063305)
19. Krause DJ, Whitaker RJ. 2015 Inferring speciation processes from patterns of natural variation in microbial genomes. *Syst. Biol.* **64**, 926–935. (doi:10.1093/sysbio/syv050)
20. Didelot X *et al.* 2011 Recombination and Population Structure in Salmonella enterica. *PLoS Genet.* **7**, e1002191. (doi:10.1371/journal.pgen.1002191)
21. Sheppard SK *et al.* 2013 Progressive genome-wide introgression in agricultural Campylobacter coli. *Mol. Ecol.* **22**, 1051–1064. (doi:10.1111/mec.12162)
22. Sheppard SK *et al.* 2013 Genome-wide association study identifies vitamin B5 biosynthesis as a host specificity factor in Campylobacter. *Proc Natl Acad Sci USA* **110**, 11923–7. (doi:10.5061/dryad.28n35.)
23. Hedge J, Wilson DJ. 2016 Practical Approaches for Detecting Selection in Microbial Genomes. *PLOS Comput. Biol.* **12**, e1004739. (doi:10.1371/journal.pcbi.1004739)
24. Chewapreecha C *et al.* 2014 Dense genomic sampling identifies highways of pneumococcal recombination. *Nat. Genet.* **46**, 305–309. (doi:10.1038/ng.2895)
25. Croucher NJ *et al.* 2011 Rapid pneumococcal evolution in response to clinical interventions. *Science* **331**, 430–434. (doi:10.1126/science.1198545)

26. Didelot X, Croucher NJ, Bentley SD, Harris SR, Wilson DJ. 2018 Bayesian inference of ancestral dates on bacterial phylogenetic trees. *Nucleic Acids Res.* **46**, e134. (doi:10.1093/nar/gky783)
27. Collins C, Didelot X. 2018 A phylogenetic method to perform genome-wide association studies in microbes that accounts for population structure and recombination. *PLoS Comput. Biol.* **14**, e1005958. (doi:10.1371/journal.pcbi.1005958)
28. Kingman JFC. 1982 The coalescent. *Stoch. Process. their Appl.* **13**, 235–248. (doi:10.1016/0304-4149(82)90011-4)
29. Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010 New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* **59**, 307–21. (doi:10.1093/sysbio/syq010)
30. Robinson DF, Foulds LR. 1981 Comparison of phylogenetic trees. *Math. Biosci.* **53**, 131–147. (doi:10.1016/0025-5564(81)90043-2)
31. Stamatakis A. 2014 RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313. (doi:10.1093/bioinformatics/btu033)
32. To T-H, Jung M, Lycett S, Gascuel O. 2016 Fast dating using least-squares criteria and algorithms. *Syst. Biol.* **65**, 82–97. (doi:10.1093/sysbio/syv068)
33. Volz EM, Frost SDW. 2017 Scalable relaxed clock phylogenetic dating. *Virus Evol.* **3**, vex025.
34. Sagulenko P, Puller V, Neher RA. 2018 TreeTime: Maximum likelihood phylodynamic analysis. *Virus Evol.* **4**, vex042. (doi:10.1101/153494)
35. Guindon S. 2010 Bayesian estimation of divergence times from large sequence alignments. *Mol. Biol. Evol.* **27**, 1768–1781. (doi:10.1093/molbev/msq060)
36. Didelot X, Siveroni I, Volz EM. 2021 Additive uncorrelated relaxed clock models for the dating of genomic epidemiology phylogenies. *Mol. Biol. Evol.* **38**, 307–317. (doi:10.1093/molbev/msaa193)
37. Meligkotsidou L, Fearnhead P. 2007 Postprocessing of genealogical trees. *Genetics* **177**, 347–358. (doi:10.1534/genetics.107.071910)
38. Didelot X, Fraser C, Gardy J, Colijn C. 2017 Genomic infectious disease epidemiology in partially sampled and ongoing outbreaks. *Mol. Biol. Evol.* **34**, 997–1007. (doi:10.1093/molbev/msw275)
39. Duchêne S, Holt KE, Weill F-X, Le Hello S, Hawkey J, Edwards DJ, Fourment M, Holmes EC. 2016 Genome-scale rates of evolutionary change in bacteria. *Microb. Genomics* **2**, e000094. (doi:10.1101/069492)
40. Duchene S, Lemey P, Stadler T, Ho SYW, Duchene DA, Dhanasekaran V, Baele G. 2020 Bayesian evaluation of temporal signal in measurably evolving populations. *Mol. Biol. Evol.* **37**, 3363–3379. (doi:10.1093/molbev/msaa163)
41. Duchêne S, Duchêne D, Holmes EC, Ho SYW. 2015 The performance of the date-randomization test in phylogenetic analyses of time-structured virus data. *Mol. Biol. Evol.* **32**, 1895–1906. (doi:10.1093/molbev/msv056)
42. Tarrío R, Rodríguez-Trelles F, Ayala FJ. 2000 Tree rooting with outgroups when they differ in their nucleotide composition from the ingroup: The Drosophila saltans and willistoni groups, a case study. *Mol. Phylogenet. Evol.* **16**, 344–349. (doi:10.1006/mpev.2000.0813)
43. Hess PN, De Moraes Russo CA. 2007 An empirical test of the midpoint rooting method. *Biol. J. Linn. Soc.* **92**, 669–674. (doi:10.1111/j.1095-8312.2007.00864.x)
44. Achtman M. 2016 How old are bacterial pathogens? *Proc. R. Soc. B Biol. Sci.* **283**, 20160990. (doi:10.1098/rspb.2016.0990)
45. Ward MJ, Gibbons CL, McAdam PR, van Bunnik BAD, Girvan EK, Edwards GF, Fitzgerald JR, Woolhouse MEJ. 2014 Time-scaled evolutionary analysis of the transmission and antibiotic resistance dynamics of Staphylococcus aureus clonal complex 398. *Appl. Environ. Microbiol.* **80**, 7275–7282. (doi:10.1128/AEM.01777-14)
46. Holden MTG *et al.* 2013 A genomic portrait of the emergence, evolution and global spread of a methicillin resistant Staphylococcus aureus pandemic. *Genome Res* **23**, 653–64.
47. Didelot X *et al.* 2012 Microevolutionary analysis of Clostridium difficile genomes to investigate transmission. *Genome Biol.* **13**, R118. (doi:10.1186/gb-2012-13-12-r118)
48. Cheng L, Connor TR, Sirén J, Aanensen DM, Corander J. 2013 Hierarchical and spatially explicit clustering of DNA sequences with BAPS software. *Mol. Biol. Evol.* **30**, 1224–1228. (doi:10.1093/molbev/mst028)
49. Tonkin-Hill G, Lees JA, Bentley SD, Frost SDW, Corander J. 2019 Fast hierarchical Bayesian analysis of population structure. *Nucleic Acids Res.* , 1–11. (doi:10.1093/nar/gkz361)
50. Lees JA, Harris SR, Tonkin-Hill G, Gladstone RA, Lo SW, Weiser JN, Corander J, Bentley SD, Croucher NJ. 2019 Fast and flexible bacterial genomic epidemiology with PopPUNK. *Genome Res.* **29**, 304–316. (doi:10.1101/gr.241455.118)
51. Hunt DEDE, David LA, Gevers D, Preheim SP, Alm EJ, Polz MF. 2008 Resource partitioning and sympatric differentiation among closely related bacterioplankton. *Science* **1081**, 1081. (doi:10.1126/science.1157890)
52. Ansari MA, Didelot X. 2016 Bayesian Inference of the Evolution of a Phenotype Distribution on a Phylogenetic Tree. *Genetics* **204**, 89–98. (doi:10.1534/genetics.116.190496)
53. Behr M, Ansari MA, Munk A, Holmes C. 2019 Testing for dependence on tree structures. *bioRxiv* , 622811. (doi:10.1101/622811)
54. Volz EM, Wiuf C, Grad YH, Frost SDW, Dennis AM, Didelot X. 2020 Identification of hidden population structure in time-scaled phylogenies. *Syst. Biol.* **69**, 884–896. (doi:10.1093/sysbio/syaa009)

55. Helekal D, Ledda A, Volz E, Wyllie D, Didelot X. 2021 Bayesian inference of clonal expansions in a dated phylogeny. *bioRxiv*, 10.1101/2021.07.01.450370. (doi:10.1101/2021.07.01.450370)
56. Ho SYW, Shapiro B. 2011 Skyline-plot methods for estimating demographic history from nucleotide sequences. *Mol. Ecol. Resour.* **11**, 423–434. (doi:10.1111/j.1755-0998.2011.02988.x)
57. Lan S, Palacios JA, Karcher M, Minin VN, Shahbaba B. 2015 An efficient Bayesian inference framework for coalescent-based nonparametric phylodynamics. *Bioinformatics* **31**, 3282–3289. (doi:10.1093/bioinformatics/btv378)
58. Karcher MD, Palacios JA, Lan S, Minin VN. 2017 phylodyn: an R package for phylodynamic simulation and inference. *Mol. Ecol. Resour.* **17**, 96–100. (doi:10.1111/1755-0998.12630)
59. Volz EM, Didelot X. 2018 Modeling the Growth and Decline of Pathogen Effective Population Size Provides Insight into Epidemic Dynamics and Drivers of Antimicrobial Resistance. *Syst. Biol.* **67**, 719–728. (doi:10.1093/sysbio/syy007)
60. Didelot X, Volz E. 2021 Maximum likelihood inference of pathogen population size history from a phylogeny. *bioRxiv*, 427056. (doi:10.1101/2021.01.18.427056)
61. Tang L, Zhou Y, Wang L, Purkayastha S, Zhang L, He J, Wang F, Song PXK. 2020 A Review of Multi-Compartment Infectious Disease Models. *Int. Stat. Rev.* **88**, 462–513. (doi:10.1111/insr.12402)
62. Volz EM. 2012 Complex population dynamics and the coalescent under neutrality. *Genetics* **190**, 187–201. (doi:10.1534/genetics.111.134627)
63. Volz EM, Koelle K, Bedford T. 2013 Viral Phylodynamics. *PLoS Comput. Biol.* **9**, e1002947. (doi:10.1371/journal.pcbi.1002947)
64. Volz EM, Kosakovsky Pond SL, Ward MJ, Leigh Brown AJ, Frost SDW. 2009 Phylodynamics of infectious disease epidemics. *Genetics* **183**, 1421–30. (doi:10.1534/genetics.109.106021)
65. Volz EM, Siveroni I. 2018 Bayesian phylodynamic inference with complex models. *PLOS Comput. Biol.* **14**, e1006546. (doi:10.1371/journal.pcbi.1006546)
66. Paradis E, Schliep K. 2019 Ape 5.0: An environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* **35**, 526–528. (doi:10.1093/bioinformatics/bty633)
67. Hadfield J, Megill C, Bell SM, Huddleston J, Potter B, Callender C, Sagulenko P, Bedford T, Neher RA. 2018 NextStrain: Real-time tracking of pathogen evolution. *Bioinformatics* **34**, 4121–4123. (doi:10.1093/bioinformatics/bty407)
68. Mutreja A *et al.* 2011 Evidence for several waves of global transmission in the seventh cholera pandemic. *Nature* **477**, 462–465. (doi:10.1038/nature10392)
69. Didelot X, Pang B, Zhou Z, McCann A, Ni P, Li D, Achtman M, Kan B. 2015 The Role of China in the Global Spread of the Current Cholera Pandemic. *PLoS Genet.* **11**, e1005072. (doi:10.1371/journal.pgen.1005072)
70. Jombart T, Eggo RM, Dodd PJ, Balloux F. 2011 Reconstructing disease outbreaks from genetic data: a graph approach. *Heredity (Edinb).* **106**, 383–90. (doi:10.1038/hdy.2010.78)
71. Didelot X, Walker AS, Peto TE, Crook DW, Wilson DJ. 2016 Within-host evolution of bacterial pathogens. *Nat. Rev. Microbiol.* **14**, 150–162. (doi:10.1038/nrmicro.2015.13)
72. Didelot X, Gardy J, Colijn C. 2014 Bayesian inference of infectious disease transmission from whole genome sequence data. *Mol. Biol. Evol.* **31**, 1869–1879. (doi:10.1093/molbev/msu121)
73. Didelot X, Kendall M, Xu Y, White PJ, McCarthy N. 2021 Genomic Epidemiology Analysis of Infectious Disease Outbreaks Using TransPhylo. *Curr. Protoc.* **1**, 1–23. (doi:10.1002/cpz1.60)
74. Biek R, Pybus OG, Lloyd-Smith JO, Didelot X. 2015 Measurably evolving pathogens in the genomic era. *Trends Ecol. Evol.* **30**, 306–313. (doi:10.1016/j.tree.2015.03.009)
75. Nylander JAA, Olsson U, Alström P, Sanmartín I. 2008 Accounting for phylogenetic uncertainty in biogeography: a Bayesian approach to dispersal-vicariance analysis of the thrushes (Aves: Turdus). *Syst. Biol.* **57**, 257–68. (doi:10.1080/10635150802044003)
76. Harris SRR *et al.* 2010 Evolution of MRSA During Hospital Transmission and Intercontinental Spread. *Science* **327**, 469–474. (doi:10.1126/science.1182395)
77. Castillo-Ramírez S *et al.* 2012 Phylogeographic variation in recombination rates within a global clone of methicillin-resistant Staphylococcus aureus. *Genome Biol.* **13**, R126. (doi:10.1186/gb-2012-13-12-r126)
78. Hsu LY *et al.* 2015 Evolutionary dynamics of methicillin-resistant Staphylococcus aureus within a healthcare system. *Genome Biol.* **16**, 1–13. (doi:10.1186/s13059-015-0643-z)
79. Tong SYC *et al.* 2015 Genome sequencing defines phylogeny and spread of methicillin-resistant Staphylococcus aureus in a high transmission setting. *Genome Res.* **25**, 111–118. (doi:10.1101/gr.174730.114.Freely)
80. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL. 2004 Versatile and open software for comparing large genomes. *Genome Biol.* **5**, R12. (doi:10.1186/gb-2004-5-2-r12)
81. Holden MTG *et al.* 2010 Genome sequence of a recently emerged, highly transmissible, multi-antibiotic- and antiseptic-resistant variant of methicillin-resistant Staphylococcus aureus, sequence type 239 (TW). *J. Bacteriol.* **192**, 888–92. (doi:10.1128/JB.01255-09)
82. Li H, Durbin R. 2009 Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760.
83. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009 The Sequence Alignment/Map (SAM) Format and SAMtools. *Bioinformatics* **25**, 2078–2079. (doi:10.1093/bioinformatics/btp352)
84. Méric G *et al.* 2015 Ecological overlap and horizontal gene transfer in Staphylococcus aureus and Staphylococcus epidermidis. *Genome Biol. Evol.* **7**, 1313–1328. (doi:10.1093/gbe/evv066)

85. Everitt RG *et al.* 2014 Mobile elements drive recombination hotspots in the core genome of Staphylococcus aureus. *Nat. Commun.* **5**, 3956. (doi:10.1038/ncomms4956)
86. Richardson EJ *et al.* 2018 Gene exchange drives the ecological success of a multi-host bacterial pathogen. *Nat. Ecol. Evol.* **2**, 1468–1478. (doi:10.1038/s41559-018-0617-0)
87. Vos M, Didelot X. 2009 A comparison of homologous recombination rates in bacteria and archaea. *ISME J.* **3**, 199–208. (doi:10.1038/ismej.2008.93)
88. Yahara K, Didelot X, Jolley KA, Kobayashi I, Maiden MCJ, Sheppard SK, Falush D. 2016 The landscape of realized homologous recombination in pathogenic bacteria. *Mol. Biol. Evol.* **33**, 456–471. (doi:10.1093/molbev/msv237)
89. Baines SL *et al.* 2015 Convergent adaptation in the dominant global hospital clone ST239 of methicillin-resistant Staphylococcus aureus. *MBio* **6**. (doi:10.1128/mBio.00080-15)
90. Jacob PE, Murray LM, Holmes CC, Robert CP. 2017 Better together? Statistical learning in models made of modules. , 1–31.
91. Novák Á, Miklós I, Lyngsø R, Hein J. 2008 StatAlign: An extendable software package for joint Bayesian estimation of alignments and evolutionary trees. *Bioinformatics* **24**, 2403–2404. (doi:10.1093/bioinformatics/btn457)
92. Herman JL, Challis CJ, Novák Á, Hein J, Schmidler SC. 2014 Simultaneous Bayesian estimation of alignment and phylogeny under a joint model of protein sequence and structure. *Mol. Biol. Evol.* **31**, 2251–2266. (doi:10.1093/molbev/msu184)
93. Duchene S, Duchene DA, Geoghegan JL, Dyson ZA, Hawkey J, Holt KE. 2018 Inferring demographic parameters in bacterial genomic data using Bayesian and hybrid phylogenetic methods. *BMC Evol. Biol.* **18**, 1–11. (doi:10.1186/s12862-018-1210-5)

Figure legends

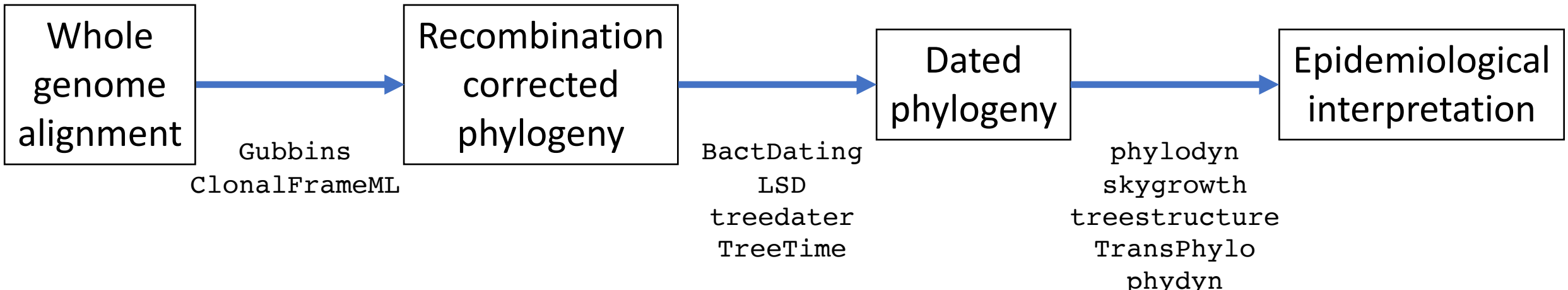
Figure 1: Overview of the step-by-step analytical approach. The names of some of the software tools that can be used in each step are indicated under the arrows.

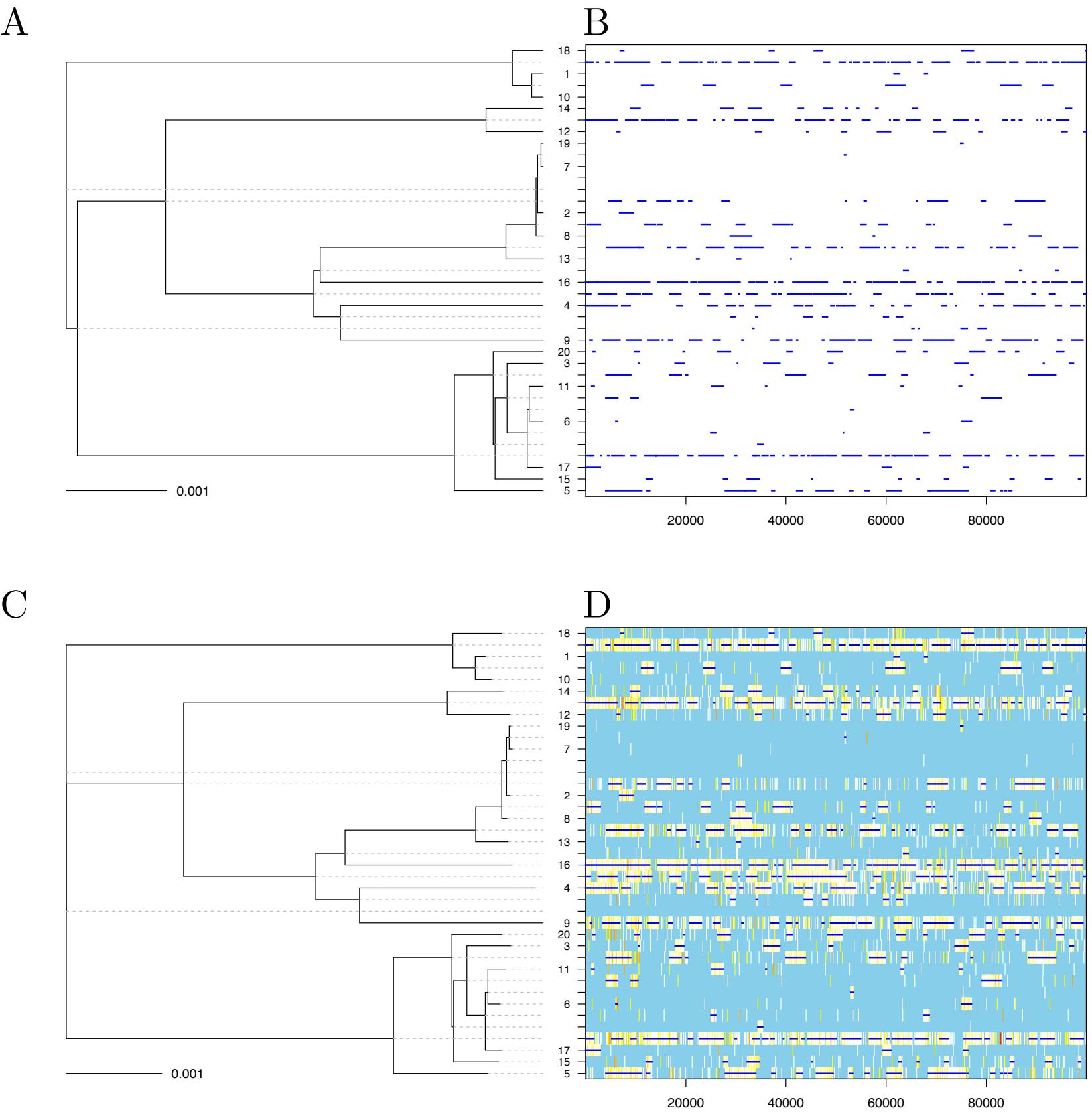
Figure 2: Illustration of the effect of recombination on phylogenetic inference. A phylogeny was simulated (A) with recombination events happening on the branches at a constant rate (B). ClonalFrameML was applied to this simulated dataset, resulting in a good reconstruction of both the clonal genealogy (C) and recombination events (D).

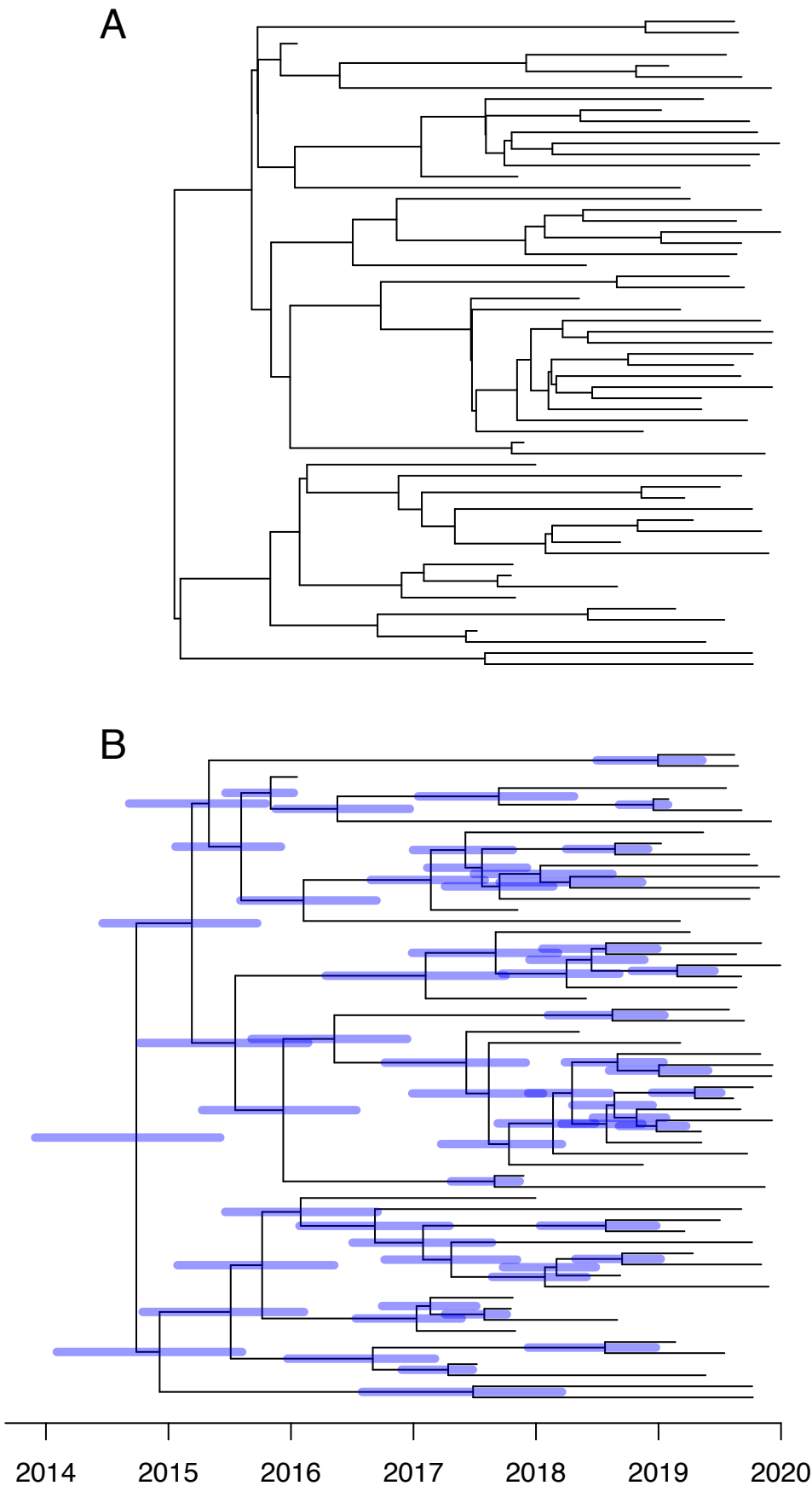
Figure 3: Illustration of the relative lack of effect of the prior model used for the inference of dated phylogeny. A dated phylogeny (A) was simulated from an epidemic model and dating was inferred (B) based on a coalescent model with constant population size.

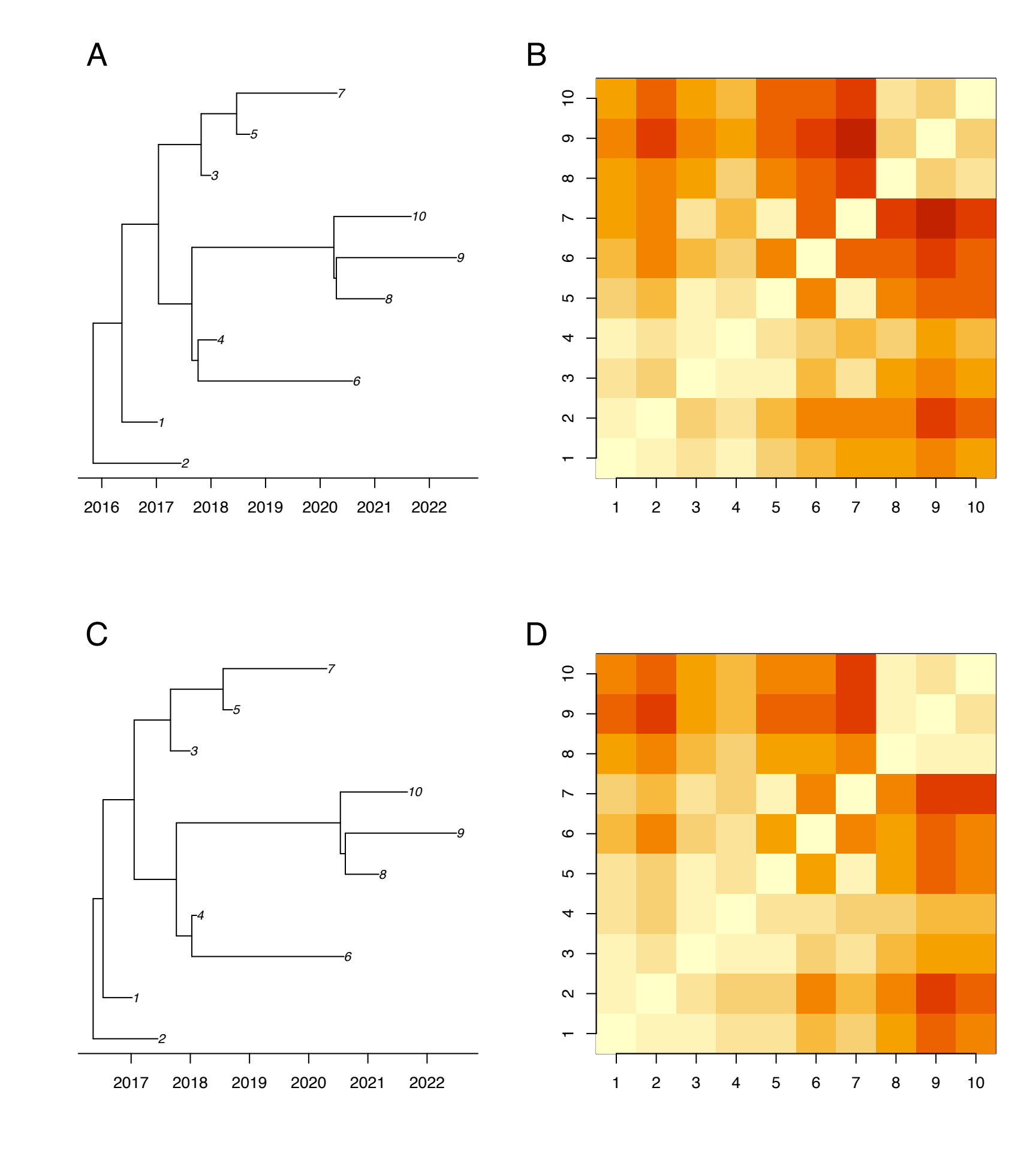
Figure 4: Illustration of the relative lack of effect of the uncertainty in the reconstructed dated phylogeny on interpretation as a tree transmission trees. Two dated phylogenies were sampled from the posterior (A and C) and a separate inference of the transmission tree was performed for each one (B and D). The coloured matrices represent the distance between pairs of cases in number of transmission links, with red coding for direct transmission and yellow coding for a distance of ten links.

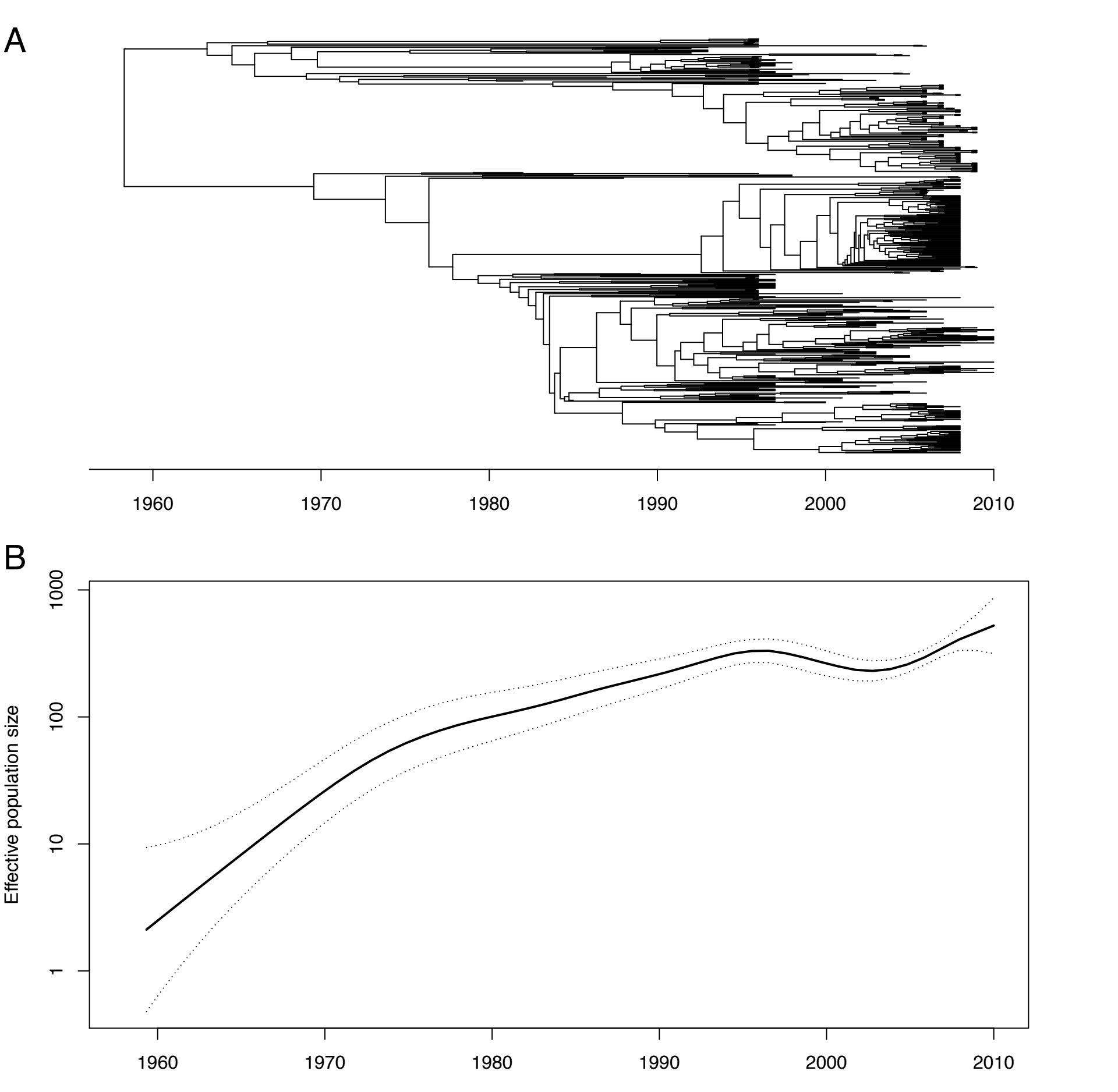
Figure 5: Example of application of the approach to a collection of *Staphylococcus aureus* ST239 genomes. The dated phylogeny was inferred using BactDating (A) and the past population size dynamics was inferred using skygrowth (B).


Whole genome alignment
Gubbins
ClonalFrameML
Recombination corrected phylogeny
BactDating
LSD
treedater
TreeTime
Dated phylogeny
phylodyn
skygrowth
treestructure
TransPhylo
phydyn
Epidemiological interpretation


A
B
C
D
18
1
10
14
12
19
7
2
8
13
16
4
9
20
3
11
6
17
15
5
0.001
20000
40000
60000
80000


A
B
2014
2015
2016
2017
2018
2019
2020


A
7
5
3
10
9
8
4
6
1
2
2016
2017
2018
2019
2020
2021
2022
B
10
9
8
7
6
5
4
3
2
1
1
2
3
4
5
6
7
8
9
10
C
7
5
3
10
9
8
4
6
1
2
2017
2018
2019
2020
2021
2022
D
10
9
8
7
6
5
4
3
2
1
1
2
3
4
5
6
7
8
9
10


A
1960
1970
1980
1990
2000
2010
B
Effective population size
1000
100
10
1
1960
1970
1980
1990
2000
2010