Check for updates

RESEARCH ARTICLE

REVISED Evaluation of phylogenetic reconstruction methods using bacterial whole genomes: a simulation based study [version 2; referees: 2 approved, 1 approved with reservations]

John A. Lees [1,2], Michelle Kendall [3], Julian Parkhill [1], Caroline Colijn [3], Stephen D. Bentley[1], Simon R. Harris [1]

[1]Infection Genomics, Wellcome Sanger Institute, Hinxton, Cambridgeshire, CB10 1SA, UK
[2]Department of Microbiology, New York School of Medicine, New York, 10016, USA
[3]Department of Mathematics, Imperial College London, London, SW7 2AZ, UK

v2 **First published:** 23 Mar 2018, **3**:33 (doi: 10.12688/wellcomeopenres.14265.1)
Latest published: 29 May 2018, **3**:33 (doi: 10.12688/wellcomeopenres.14265.2)

Abstract

Background: Phylogenetic reconstruction is a necessary first step in many analyses which use whole genome sequence data from bacterial populations. There are many available methods to infer phylogenies, and these have various advantages and disadvantages, but few unbiased comparisons of the range of approaches have been made.
Methods: We simulated data from a defined 'true tree' using a realistic evolutionary model. We built phylogenies from this data using a range of methods, and compared reconstructed trees to the true tree using two measures, noting the computational time needed for different phylogenetic reconstructions. We also used real data from *Streptococcus pneumoniae* alignments to compare individual core gene trees to a core genome tree.
Results: We found that, as expected, maximum likelihood trees from good quality alignments were the most accurate, but also the most computationally intensive. Using less accurate phylogenetic reconstruction methods, we were able to obtain results of comparable accuracy; we found that approximate results can rapidly be obtained using genetic distance based methods. In real data we found that highly conserved core genes, such as those involved in translation, gave an inaccurate tree topology, whereas genes involved in recombination events gave inaccurate branch lengths. We also show a tree-of-trees, relating the results of different phylogenetic reconstructions to each other.
Conclusions: We recommend three approaches, depending on requirements for accuracy and computational time. For the most accurate tree, use of either RAxML or IQ-TREE with an alignment of variable sites produced by mapping to a reference genome is best. Quicker approaches that do not perform full maximum likelihood optimisation may be useful for many analyses requiring a phylogeny, as generating a high quality input alignment is likely to be the major limiting factor of accurate tree topology. We have publicly released our simulated data and code to enable further comparisons.

Keywords

phylogeny, simulation, tree distance, bacteria, phylogenetic methods

Open Peer Review

Referee Status: ✓ ✓ ?

	Invited Referees 1	 2	 3
REVISED **version 2** published 29 May 2018	✓ report		
version 1 published 23 Mar 2018	? report	✓ report	? report

1 **Lauren A. Cowley**, Harvard T.H. Chan School of Public Health, USA
Taj Azarian , Harvard University, USA

2 **Philip M. Ashton** , Oxford University Clinical Research Unit, Vietnam

3 **João A. Carriço** , University of Lisbon, Portugal

Discuss this article

Comments (1)

Corresponding authors: John A. Lees (john.lees@nyumc.org), Simon R. Harris (sh16@sanger.ac.uk)

Author roles: Lees JA: Conceptualization, Formal Analysis, Investigation, Methodology, Software, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Kendall M**: Formal Analysis, Methodology, Software, Visualization, Writing – Review & Editing; **Parkhill J**: Funding Acquisition, Resources, Supervision, Writing – Review & Editing; **Colijn C**: Methodology, Resources, Supervision, Writing – Review & Editing; **Bentley SD**: Resources, Supervision, Writing – Review & Editing; **Harris SR**: Conceptualization, Methodology, Supervision, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

How to cite this article: Lees JA, Kendall M, Parkhill J *et al.* **Evaluation of phylogenetic reconstruction methods using bacterial whole genomes: a simulation based study [version 2; referees: 2 approved, 1 approved with reservations]** Wellcome Open Research 2018, **3**:33 (doi: 10.12688/wellcomeopenres.14265.2)

Grant information: This work was supported by the Wellcome Trust (098051). JAL was also supported by a Medical Research Council studentship grant (1365620). CC was supported by the Engineering and Physical Sciences Research Council EPSRC EP/K026003/1 and EPSRC EP/N014529/1.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

First published: 23 Mar 2018, **3**:33 (doi: 10.12688/wellcomeopenres.14265.1)

REVISED **Amendments from Version 1**

In this version we have made suggested additions and changes from the four reviewers.

New analysis includes the addition of two new tree building methods into the comparisons: a neighbour-joining tree from Hamming distances, and a maximum likelihood tree using both SNP and binary gene presence/absence as input partitions. We have also included a comparison of using ascertainment bias correction in the two best performing methods, as suggested by a reviewer. On the simulated data we looked at the treespace of a posterior of trees from a Bayesian method, to further explore uncertainty output by phylogenetic software methods. Using real data, we looked at the accuracy of a tree from all ribosomal proteins, and found it to be more accurate than any single gene tree. We have updated Figure 4, and the Supplementary materials.

See referee reports

Introduction

Phylogenetic analysis is a complex task, but one that is foundational to many applications in bacterial genetics: molecular evolution, outbreak tracing and genomic epidemiology, to name a few[1,2]. The modern genomic analyst faces a bewildering array of options at every stage of the process.

The possible number of trees for even a small number of tips is enormous[3] – for 96 tips there are 10^{173} possible trees (compare this to 10^{80} atoms in the observable Universe, or even 10^{120} possible games of chess). Fortunately, sophisticated software methods allow us to sensibly navigate through this space to the most likely trees.

Generally the steps taken when analysing a population of bacteria that have been whole genome sequenced are as follows. Quality control of the raw data must first be performed, after which a whole-genome alignment of the sequences is produced. The alignment is usually produced by mapping reads to a reference sequence (of which many likely exist), but may also be obtained by *de novo* assembly followed by whole-genome alignment (either by progressive local alignment, or through multiple sequence alignment of orthologous genes and intergenic regions). Many methods are available to map reads to a reference, assemble reads into contigs and align contigs or genes, and each method will typically have many options. This alignment is the key input for phylogenetic inference software. Even more methods, with yet more complex options, exist to determine the most likely phylogeny given a sequence alignment. Alternatively, one may forgo alignment altogether, and opt instead for a k-mer distance-based approach followed by a neighbor joining tree.

Understandably, this complexity and range of choice means that methods sections of papers using phylogenetic analysis are often different between studies. This disparity is likely due to different software preferences (familiarity, speed and usability being major factors in this choice), rather than an informed choice based on the biological question and resources to hand. One should carefully consider what question the tree is trying to address: is it to look at overall population structure, or to try and find precise relationships between closely related isolates? The relative merits of different approaches are difficult to objectively assess, even after careful reading of the original method manuscripts. The potential effect of different combinations of approaches at each step in the process between raw sequence reads and the final phylogeny has seldom been explored.

It is therefore desirable to provide a comparison between phylogenetic methods that is focused on methods' ability to answer the biological question at hand. Some previous attempts have been made, using either simulated data, experimental evolution, or an assumption that the maximum likelihood phylogeny is correct. One such study assessed the running times and likelihood of trees drawn from simulated data using two pieces of software (RAxML and FastTree), assuming the model of sequence evolution is correct[4]. A larger study in eukaryotes compared these two methods with IQ-TREE in terms of the best likelihood obtained using both species and gene trees[5]. Other small-scale comparisons include a comparison of read-to-tree pipelines with other pieces of software[6], and the production of "well characterised" reference datasets for testing methods[7]. A recent study instead used an *Escherichia coli* hypermutator to conduct experimental evolution along a defined balanced phylogeny, and then by sequencing the strains at the tips, the authors compared the ability of 12 combinations of methods to reconstruct the correct phylogenetic relationship[8]. An overview of how the most commonly used combinations of methods perform in terms of phylogeny accuracy, as opposed to best likelihood, does not yet exist. Comparison of likelihoods alone assumes that we know the true evolutionary model, and doesn't allow us to evaluate in what way the tree is wrong.

In this paper we present a simulation-based analysis of the speed, ease of use, and accuracy of some of the common ways to obtain a phylogeny from bacterial whole genome sequence data. We define a true tree, from which we produce whole genome sequence data using realistic simulations (thereby avoiding the problem of circularity of model choice). A range of methods are then evaluated for accuracy using appropriate metrics in tree space. We hope to provide some insight into which approaches should be favoured in certain settings while acknowledging that our simulations are far from comprehensive. We also make our code and simulated data publicly available in the hope that this might inspire further method comparisons aimed at different settings.

Methods

Simulating bacterial populations – assemblies and alignments

We wished to simulate genomes in a realistic way, without using the same model of evolution that any one software package uses to compute tree likelihoods or sequence distances in order to reconstruct the tree. This would be circular, and would result in that software package necessarily performing best.

For the simulations we used parameters for *Streptococcus pneumoniae*, whose evolution has been extensively studied using genomic data, but artificially used a tree topology from another species which had desirable properties for downstream comparisons. We therefore used Artificial Life Framework v1.0 (ALF)[9] to simulate evolution along a given phylogenetic tree, for the 2 232 coding sequences in the *S. pneumoniae* ATCC 700669 genome[10] as the MRCA. As well as modeling SNP evolution, ALF also allows for short insertions and deletions (INDELs), gene loss and horizontal gene transfer events which occur in real populations but are usually not included in phylogenetic models. In parallel, we used DAWG v1.2[11] to simulate evolution of intergenic regions (defined as sequence not annotated as a CDS). We identified a phylogeny (Figure 1), originally produced by Kremer *et al.*[12] from a core genome alignment of 96 *Listeria monocytogenes* genomes from patients with bacterial meningitis which possessed a number of qualities we wished to be able to reproduce. Particularly, it had two distinct lineages (also making midpoint rooting suitable, and negating the strong dependence on correct rooting implicit in the Kendall and Colijn metric[13]), several clonal groups within each lineage, long branches and a polyphyletic population cluster. Population clusters were estimated from the resulting core genome alignment from simulations using Bayesian Analysis of Population Structure v6.0 (BAPS)[14]. We define *N* as the number of strains in the study and *M* as the number of aligned sites.

We used realistic parameters, as far as possible, for the simulation run with ALF. To estimate rates to use in the generalised time-reversible (GTR) matrix and the size distribution of INDELs, we first aligned *S. pneumoniae* strains R6 (AE007317), 19F (CP000921) and *Streptococcus mitis* B6 (FN568063) using Progressive Cactus v0.0[15]. This whole genome alignment allowed calculation of SNP and INDEL rates for these models. We used previously determined parameters for the rate of codon evolution[16], relative rate of SNPs to indels in coding regions[17], rates of gene loss and horizontal gene transfer[18] when running the simulation. We then used ALF with these parameters to simulate the evolution of coding sequences from the root genome along the given phylogeny. For the intergenic regions we used the same GTR matrix parameters and previously estimated intergenic SNP to INDEL rate[17]. We combined the resulting sequences of coding and non-coding regions at tips of the phylogeny while accounting for gene loss and transfer, and

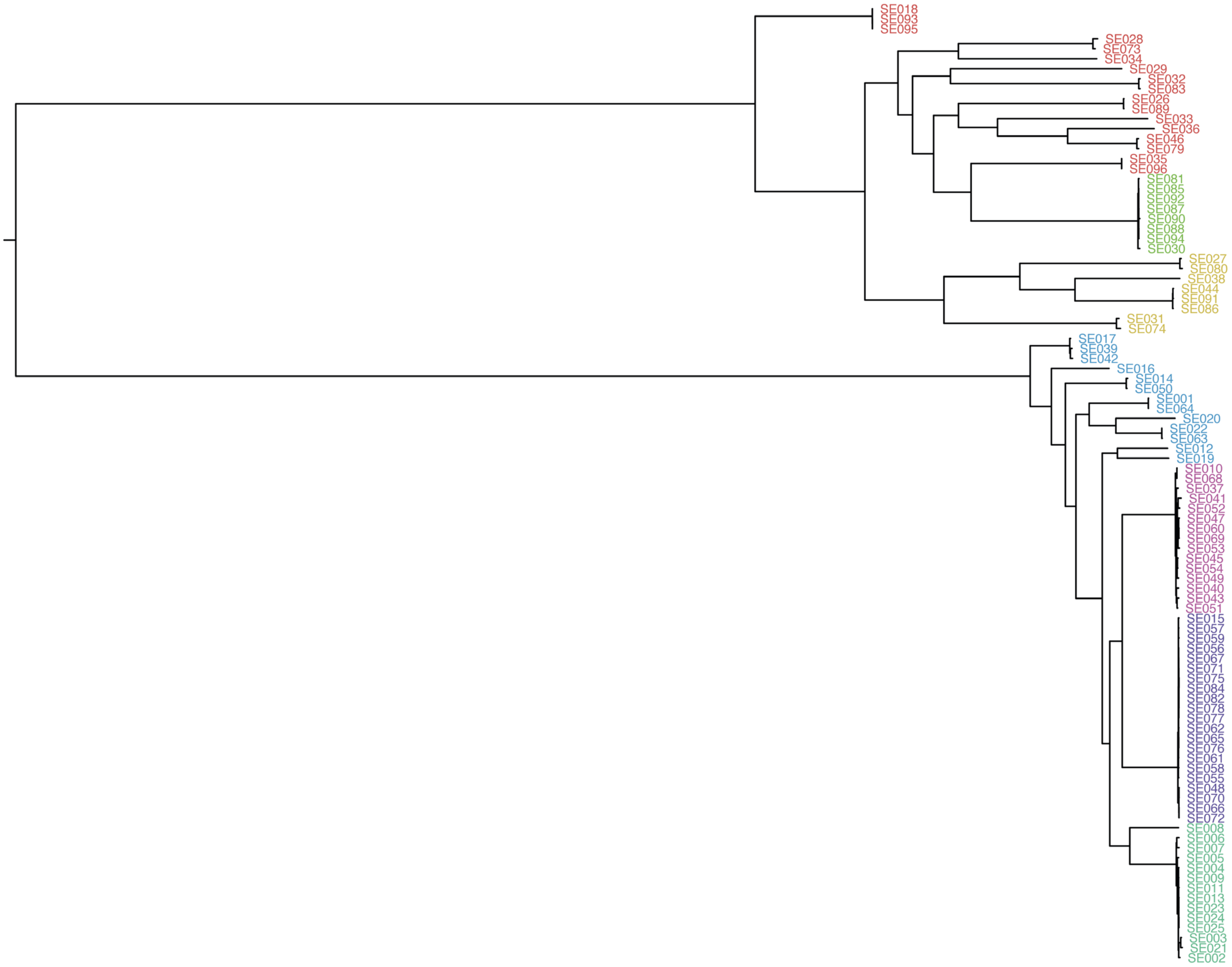

Figure 1. The phylogeny inferred by Kremer *et al.*[12] used as the true tree in simulations. Tips are coloured by BAPS cluster inferred from the core genome alignment.

finally generated error prone Illumina reads from these sequences using pIRS v1.11[19]. An overview of this process is shown in Supplementary Figure 1 (Supplementary file 1).

To generate input to phylogenetic inference algorithms, we created assemblies and alignments from the simulated reads. We assembled the simulated reads into contigs with velvet v1.2.09[20] using https://github.com/tseemann/VelvetOptimiser to choose an optimal coverage cutoff and k-mer size (between 37 and 81). We then improved and annotated the resulting scaffolds using the sanger-pathogens improvement pipeline with default parameters[21]. We generated alignments by mapping reads to the TIGR4 reference using bwa-mem v0.7.10 with default settings[22], and called variants from these alignments using samtools v1.2 mpileup and bcftools call[23]. We used Roary 1.007001[24] with a 95% BLAST ID cutoff to construct a pan-genome from the annotated assemblies, from which a core gene alignment was created with MAFFT v7.205[25]. Downstream analysis using genes was done using this pan-genome. We then created alignments using two further methods. For an MLST-like alignment we selected seven genes at random from the core alignment (present in all strains) which had not been involved in horizontal transfer events. For a Progressive Cactus alignment, we ran the software on the assemblies using default settings, and extracted regions aligned between all genomes from the hierarchical alignment file and concatenated them.

Methods of phylogeny reconstruction

Using the nucleotide alignments described above as input, we ran the following phylogenetic inference methods:

- RAxML v7.8.6[26] with a GTR+gamma model (-m GTRGAMMA).
- RAxML v7.8.6 with a binary+gamma sites model (-m BINGAMMA).
- IQ-TREE v1.6.beta4[27] using a GTR+gamma model (-m GTR+G) (denoted slow) and using GTR and the -fast option (denoted fast).
- IQ-TREE v1.6.beta4 with mixed partitions with matched branch lengths and varying evolutionary rates (-spp). We used a GTR+gamma model (-m GTR+G) for the SNP alignment, and a binary GTR model (-m GTR2) for gene presence/absence.
- FastTree v2.1.9[28] using the GTR model (denoted slow) and using the -pseudo and -fastest options (denoted fast).
- Parsnp v1.2[29] on all assemblies using the -c and -x options (removing recombination with PhiPack).

We attempted to run the REALPHY v1.12 pipeline[6], but it was not computationally feasible due to the slow mapping step (using bowtie2) not being parallelisable by strain.

We also created pairwise distance matrices using:

- Mash v1.0[30] (default settings) between assemblies.
- Andi v0.9.2[31] (default settings) between assemblies.
- Hamming distance between informative k-mers using a subsample of 1% of counted k-mers from assemblies[32].
- Hamming distance between SNP sites produced by Disty McMatrixface v0.1.0.
- JC and logdet distances between sequences in the alignment, as implemented in SeaView v4.0[33].
- Distances between core gene alleles (present in 100% of isolates) from the roary alignment. We added a distance of zero for each core gene with identical sequence, or added a distance of one if nonidentical, as used in the BIGSdb genome comparator module[34].
- Normalised compression distance (NCD)[35], using PPMZ as the compression tool[36].

For all the above distance matrix methods we then constructed a neighbor joining (NJ) tree, a BIONJ tree[37] using the R package ape, and an UPGMA tree using the R package phangorn. In the comparison we retained the tree building method from these three with the lowest distance from the true tree (see below).

Quantifying differences between phylogenetic tree topologies

To measure the differences in topology between the produced trees (either between the true tree and an inferred tree, or between all different inferred trees) we used two measures. As a sensitive measure of changes in topology we used the metric proposed by Kendall and Colijn[13] setting $\lambda = 0$ (ignoring branch length differences). We choose to ignore branch length differences as maximum likelihood methods (which will perform much better) will not be comparable with distance based approaches. We also decided that topology difference was more intuitive over the range of methods we tried, rather than the combination of topology and branch lengths that setting $\lambda > 0$ would give. We compared the true tree against randomly generated trees from the ape function rmtree, which randomly splits edges. After midpoint rooting this gave 286 (95% CI 276–293) as a comparison to poor topology inference. To illustrate how these numbers correspond to actual changes in topology we used the *plotTreeDiff* function from the treespace package for three representative comparisons (see interactive treespace plots or static Supplementary Figure 2–Supplementary Figure 5 (Supplementary File 1).

For trees distant from the true tree by the KC metric it was useful to test whether the tree was accurate overall and only a few clade structures were poorly resolved, or whether the tree failed to capture important clusters at all. We therefore checked the clustering of the BAPS clusters from the true alignment on each inferred tree. We did this with both the primary BAPS cluster, which separates the two main lineages, and the secondary BAPS clusters which define finer structure in the data and includes a polyphyletic cluster. For each BAPS cluster, we assessed whether tips were clustered correctly by checking whether it was still monophyletic in the inferred tree, and whether the polyphyletic cluster was still split in the same way.

Core gene trees from real data

We used a previously generated core genome alignment from 616 *S. pneumoniae* samples isolated from the nasopharynx of asymptomatically carrying children in Massachusetts[38–41]. We ran IQ-TREE on the whole alignment using a GTR model (-m GTR). We then aligned each core gene at the codon level with RevTrans v1.10[42], and then ran IQ-TREE on each nucleotide alignment using the same model. We calculated the KC metric with $\lambda = 0$ between all these pairs of trees, and used treespace to perform multidimensional scaling in two dimensions to visualise the pair-wise distances[43–45].

Results

Table 1 and Figure 2 show the results of our simulations, ranked by their KC distance from the true tree. We note that all methods except for the NCD were able to recapitulate the population clusters as defined by BAPS. Additionally, all methods found a consistent midpoint root. This is reflected by the KC metric scores which would be significantly higher if there were 'deeper' differences in the tree topologies, particularly concerning the root position.

Table 1. Accuracy and resource usage of phylogenetic reconstruction methods, ordered by KC metric score. The method lists the best combinations of all alignment with phylogenetic method, and distance matrices with phylogenetic methods. Three scores of accuracy of the phylogeny are shown; the KC metric is described in the text, the BAPS scores (the primary and secondary clusters, respectively) are a tick if the clusters are as in the true tree, otherwise which clusters are wrong (all clusters, or just the polyphyletic clusters). Parallelisability shown is that built into the software, "completely" is when every value in a distance matrix is independent so can be parallelised up to N^2 times. Accessory indicates whether accessory elements (not present in all isolates) are used in the phylogenetic inference.

Method	KC (0-286)	BAPS 1	BAPS 2	CPU time	Memory	Overheads	Parallelisability	Accessory genome?	Recommended
RAxML + close reference alignment	4.63	✓	✓	806.5 minutes	2.7 Gb	Mapped alignment	Pthreads	No	NA (artificial)
RAxML + alignment	11.2	✓	✓	587 minutes	3.0 Gb	Mapped alignment	Pthreads	No	Accurate but slow
IQ-TREE (slow) + alignment	11.2	✓	✓	703 minutes	3.2 Gb	Mapped alignment	Pthreads or MPI	No	Accurate but slow
IQ-TREE (fast) + alignment	11.3	✓	✓	14.6 minutes	1.1 Gb	Mapped alignment	Pthreads or MPI	No	Accurate/fast tradeoff
Parsnp	14.0	✓	✓	42.5 minutes	2.6 Gb	Assemblies	Threads	No	Artificial
FastTree + alignment	16.0	✓	✓	189 minutes	10.6 Gb	Mapped alignment	Threads (up to 4)	No	Accurate/fast tradeoff
RAxML + core gene alignment	18.6	✓	✓	29.2 minutes	154 Mb	Core gene alignment	Pthreads	No	Comparable to mapping
NJ + SNPs alignment	20.5	✓	✓	Negligible	Negligible	Mapped alignment	No	No	No
IQ-TREE + mixed partitions	24.5	✓	✓	1316 minutes	3.2Gb	Mapped alignment + accessory genes	Pthreads or MPI	Yes	No
BIONJ + mash distances	51.7	✓	✓	0.75 minutes	10 Mb	Assembly	Completely	Yes	Best, when no alignment
RAxML + Seven gene alignment (MLST-like)	62.6	✓	✓	1.4 minutes	19 Mb	Assembly	Pthreads	No	No
BIONJ + andi distances	66.0	✓	polyphyly	7.48 minutes	290 Mb	Assembly	Completely	Yes	No
RAxML + Cactus alignment	67.2	✓	✓	9 600 minutes	37.4 Gb	Assembly	Threads	No	No
RAxML + gene presence/absence	77.3	✓	polyphyly	4.28 minutes	20 Mb	Core gene alignment	Threads	Yes	No
BIONJ + k-mer distances	89.6	✓	✓	37.3 minutes	180 Mb	Assembly	Threads	Yes	No
NJ + ANI/ Hamming distances	98.1	✓	polyphyly	Negligible	230 Mb	Mapped alignment	No	No	No
BIONJ + BIGSdb-like	150	✓	polyphyly	0.48 minutes	Negligible	Assembly	Completely	No	No
UPGMA + NCD	210	✓	all	1 040 minutes	Negligible	Assembly	Completely	Yes	No

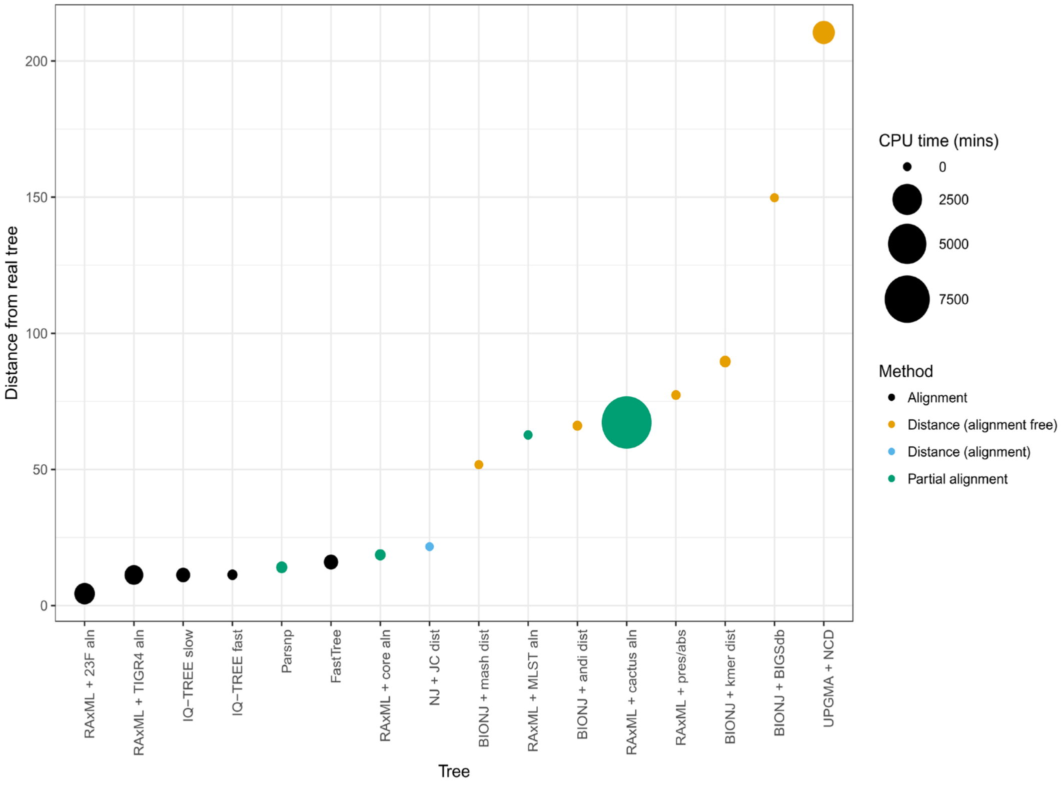

Figure 2. Ordered accuracies from Table 1, showing the CPU time required for each tree. There are large changes in accuracy between the alignment and distance methods, and again between two inaccurate distance methods.

For construction of a maximum likelihood (ML) tree, RAxML is one of the most heavily used and efficient software methods available. As expected, this was the most accurate method tested, and also the most resource heavy (apart from whole-genome alignment, discussed later). RAxML's model is a close fit to the model used to generate the data, and this model is expected to be a good model of evolution. There was no significant difference in the likelihood of the fit of the inferred tree and the true tree under this model (LRT = 2.34; p = 0.13). When using an alignment against a different reference genome from the one we actually used in the simulations, as is more likely to be the case in real alignment production, RAxML was tied for accuracy with IQ-TREE which also produced the same tree. In our simulations RAxML had better resource requirements than IQ-TREE, though over a range of data the programs are likely comparable.

A common consideration with ML trees from alignments is whether to include all sites, or remove the constant sites and analyse just SNP sites. The potential advantage of the latter approach is to reduce memory usage, which is particularly important when analysing huge alignments with thousands of sequences. Selecting just the polymorphic sites introduces an ascertainment bias which can cause branch lengths to be overestimated, so a correction needs to be applied to prevent this[46]. Both RAxML and IQ-TREE implement this correction, so we compared tree accuracy and resource use between these two modes (Supplementary Table 1; Supplementary file 1). We found similar topology in both modes, and if anything more accurate branch lengths when using polymorphic sites with an ascertainment bias correction. Most importantly, resource use (CPU time and maximum memory use) was much lower when using only variable sites – we would therefore recommend this approach over using the full alignment.

Partial alignment methods or alternative reconstruction give good trees

Knowing the quality of maximum likelihood trees, one approach a user may take to reduce the large computational requirements is to reduce the number of sites *M* that are included in the alignment. Some common ways this can be achieved are either by finding clusters of orthologous genes and only using sites from "core" genes (those present in every sample), or by using an alignment of the pre-defined MLST genes. In this test we found that using a core genome alignment slightly reduced the accuracy, whereas using an alignment of seven genes, similar to MLST, reduced the accuracy greatly, as only a small proportion of the genomic variants are now used in the inference.

Other than as a way to reduce computational burden, core genome alignment may increase the accuracy of the input alignment by excluding mismapping of repetitive regions and minimising bias from missing data in accessory genes. However, there is the issue that when a variant is present in a region overlapped by two genes it will be erroneously represented twice. When analysing a whole species, particularly when the core genome contains only a fraction of the overall diversity, this can also lead to a loss of resoution within lineages. One way to avoid this is by first defining lineages, then producing a separate alignment and tree for each. In this case one should take advantage of multiple reference genomes by selecting one that is genetically close to each lineage to produce the alignment.

When performing phylogenetic analysis, the user should consider whether they want to include the accessory genome in their inference (final column in Table 1). In this simulation, evolution of the core and accessory genome are correlated, so that including the accessory genome improves accuracy over using core genome alone. In a species such as *Streptococcus pneumoniae* where multiple distinct lineages are maintained over time, the core and accessory evolution tend to be correlated in this way[47]. In some other species, for example *Staphylococcus aureus*[48], the accessory genome is dominated by mobile elements such as transposons and phage (the same is also true within a single lineage of *S. pneumoniae*). In species such as *Escherichia coli* accessory genes are highly mobile[49]. In both cases the evolutionary signal from accessory genes is discordant from core genome evolution, so including these in the alignment will not give a good estimate of vertical evolutionary distance between strains. In other situations the core and accessory genome may both carry signals of vertical evolution, but they may be discordant with each other due to different evolutionary processes acting on each type of variation. A binary model of evolution can be used to build a maximum likelihood tree based on accessory gene gain and loss (RAxML + gene presence/absence), but we found that its accuracy is much lower than a model of SNP variation within genes. A possibility for combining these two data types would be to have separate model partitions for SNP variation and gene gain/loss. We have provided an example of this using IQ-tree on the simulated data, though we found this actually reduced accuracy of the resulting topology (KC score 24.5). Possible issues with this approach are that genes which are discordant with the phylogenetic signal from vertical evolution of the core genome (e.g. mobile genetic elements) may reduce accuracy, and incorrectly split orthologues in the accessory genome.

To further investigate core genome alignment, we compared individual gene trees to a core genome tree in a real population of *S. pneumoniae* genomes. We created trees from all core genes, and compared them by projecting pairwise KC distances into two dimensions (Figure 3). The figure shows that the core genome tree behaves like an 'average' of the individual core gene tree topologies, without being biased by the bad topologies produced at distances far from the center of the main cluster. Looking at the distant topologies, we found that the genes giving these trees were mostly ribosomal related proteins. These alignments contained very little variation due to their highly conserved function, providing little information for phylogenetic resolution – the root and ancestral part of these topologies were different from the core genome alignment tree, likely due to random placement of nodes, giving highly divergent KC distances. Reassuringly, concatenating these 82 ribosomal gene alignments and producing a tree performed better than any individual gene alignment (KC distance = 1362), giving more confidence in rMLST schemes.

The gene trees closest to the whole core gene alignment tree were those with the most variation. When we included branch lengths in the distance measure ($\lambda = 1$ in the KC metric), very short branch lengths contribute far less to the tree distance than longer lengths, and the ribosomal genes are no longer outliers. Many of the furthest gene trees from the core genome tree are from genes known to be involved in recombination events[50], as shown in Supplementary Table 2 (Supplementary File 1). Recombinations result in a large number of SNPs against a reference; because phylogenetic methods assume vertical evolution, recombination tends to inflate estimated branch lengths, but generally do not affect topology[51]. The best practice is to try to remove these regions before performing phylogenetic reconstruction[52]. When picking an MLST scheme for an organism the most important considerations are probably recapitulation of epidemiological parameters, ease and consistency of use[53]. However, given a choice of suitable genes to use, ranking of these phylogenetic signals may be a useful additional consideration. Searching through combinations of different gene alignments suggested little interaction between them affecting the final topology; the upshot being that genes that individually perform well can be considered as candidates without worrying about the specific combination chosen.

We also evaluated the quality of a phylogeny drawn from a progressiveCactus alignment[15], which performed best in a comparison between whole genome aligners[54]. Whole genome alignment uses linear sequences in an annotation-free manner, and by breaking the alignment job into smaller local regions can align sequences in the presence of structural variation such as gene gain and loss, inversions and transversions – both core and accessory elements are aligned. In this comparison, the core genome alignment we extracted was smaller than that produced by Roary, and therefore produced a less accurate phylogeny. This class of methods is therefore best suited to comparing small numbers of genomes from larger evolutionary distances (across species), rather than large numbers of more closely related genomes.

In the search for greater computational efficiency, rather than changing the alignment one may instead opt to use a different method of phylogenetic inference. One piece of software which aims to infer phylogeny faster than a maximum likelihood method, albeit at the expense of accuracy, is FastTree[28]. In our test FastTree ran four times faster than RAxML, without much decrease in accuracy. We found little difference in accuracy when using the fast and slow options. The scaling of CPU time in FastTree by number of sequences is more favourable than RAxML, so as the number of sequences increases the relative speedup of FastTree will also increase. It should also be noted that FastTree obtains around a 2x speedup from using four CPUs using OpenMP, whereas RAxML can use around 16 threads at close to 100% efficiency.

Parsnp[29] produces a core genome alignment by rapidly finding maximal exact matches (MEMs, as in nucmer) which can include both genes and intergenic regions. The use of MEMs means that assembly quality will affect parsnp results, which was designed for use with reference-quality genomes (for example, those produced by SMRT sequencing. In our test we found that it performed even better than

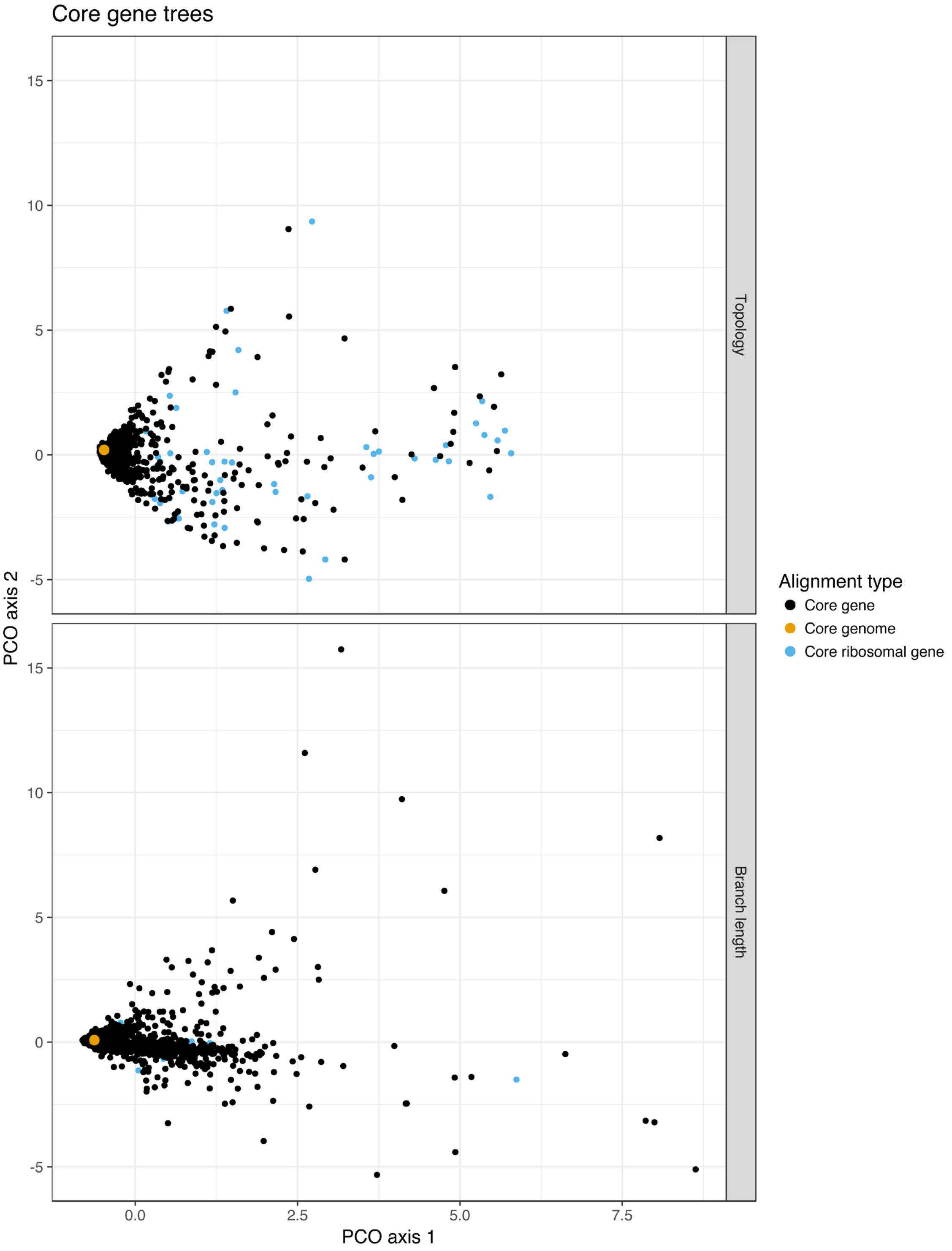

Figure 3. A multidimensional scaling plot of the KC distances between all core gene trees from a real population of 616 *S. pneumoniae* genomes. Top: topology distances ($\lambda = 0$); bottom: branch length distances ($\lambda = 0$). The core genome tree from the concatenated alignment is shown in yellow; trees from ribosomal proteins, which tended to have different topologies due to their lack of variation, are shown in blue. The top twenty divergent trees by branch length are listed in Supplementary Table 2 (Supplementary File 1). The full list of distances by gene can be accessed at https://gist.github.com/johnlees/da164a4260e13528e8315e266a46bf3f.

FastTree while using less CPU time, however our assemblies from simulation are likely more amenable to comparison of MEMs than real data, which is more fragmented. The method does not deal well with mobile elements or recombination, so extra caution should be used with real datasets where this variation is prevalent.

Finally, we saw very promising results when using the "fast" mode of IQ-TREE, currently available in beta. Reconstruction in this case was as accurate as a full maximum likelihood method, and completed quickly with modest memory requirements. Once available as a stable release, this may prove to be the most accurate way to efficiently infer large phylogenies.

Genetic distance based approaches rapidly give a rough tree topology

Early phylogenetic methods involved drawing a neighbour joining tree from a matrix of pairwise distances between all tips. This method is fast and simple. When we used distances calculated from the same alignment as RAxML this approach was somewhat worse than the reduced number of sites or reduced accuracy methods above, but still gave a good overall topology – better than an ML tree from seven core genes (similar to MLST). A tree can also be drawn from distances using BIONJ, which by using a simple evolutionary model can be expected to provide trees with more accurate topologies than NJ[37]. Another alternative is UPGMA, though as a hierarchical clustering method it would not be expected to recover the same topology as a phylogenetic method (but perhaps the same clusters).

However, in the present era, we see the main advantage of this class of methods as being able to avoid having to create an alignment from mapping[55]. If one is able to calculate genetic distances from assemblies or even directly from reads, the relatively costly and challenging step of creating a large multiple sequence alignment can be avoided. Although N^2 distances need to be evaluated, these calculations are independent so the process is trivially parallelisable. We tried creating trees from five methods which can evaluate pairwise distances rapidly: mash, andi, k-mer distances, BIGSdb and the normalised compression distance (NCD).

The NCD is a general method to compare the similarity between any two data objects[35]. The NCD between two objects x and y (in this case the sequence of assemblies) is computed as follows:

$$\mathrm{NCD}(x, y) = \frac{Z(x, y) - \min\left[Z(x), Z(y)\right]}{\max\left[Z(x), Z(y)\right]}$$

where $Z(x)$ is the size after compression of file x. The rationale is that the more two sequences are similar to each other, then the more the compression method will be able to use this similarity to reduce the overall size of the concatenated file towards the lower limit of the size of the compressed individual files. We used PPMZ as the compressor to avoid issues with minimum block size[36], but only recovered the largest scale feature of the two main lineages in the topology. This suggests the the NCD is not well suited to finding distances between sets of closely related sequences, but may perform better with more distant genomes. PPMZ may not be the best compressor overall due to its long run time, but we did not investigate this further.

BIGSdb is a database designed to store bacterial sequences, and perform pre-defined analysis rapidly on them[34]. Trees from genomes in this database can be produced with the GenomeComparator module. This works by comparing the alleles of core gene sequences, increasing the distance between two genomes by one for each allelic difference between the genes that they have. The potential advantage of this is that recombination events will correctly be counted as a single evolutionary change, rather than as multiple separate SNP differences. However, this approach also limits resolution and inference of intra-cluster distances, and produced one of the worst topologies in our tests.

Finally, we used k-mer distances[32], mash[30] and andi[31] to create distance matrices. andi counts the number of mismatches between equally spaced maximal exact matches between a pair of sequences. mash was partly designed as an improvement to the accuracy of andi, and instead uses the MinHash algorithm to rapidly approximate the Jaccard distance between the sets of k-mers in each assembly. This is also the distance approximated by our k-mer method, but is many-fold more efficient due to the use of MinHash. In our test, we found that mash performed the best out of any distance-based measure in accuracy and efficiency, but was still significantly less accurate than the alignment-based methods. Considering the ease of use and efficiency of mash, its ability to recover population clusters means that it could be recommended as the tool of choice for first-pass analysis.

Discussion

We have analysed the ability of a range of phylogenetic inference methods to reproduce the topology and clustering of a known tree when given realistic simulated data derived from the same known tree. Figure 4 shows an alternative presentation of our results: a tree-of-trees, also showing the ways in which some of the incorrect trees may be similar to each other.

Overall, we found that modern maximum likelihood methods and a good alignment can obtain an accurate phylogeny in reasonable runtimes; using approximate phylogeny methods with a good alignment is the next best thing, followed by reducing the alignment size. The best quality results had the longest computational time requirements, consistent with our mechanistic understanding of how phylogenetic inference should perform. We would expect maximum likelihood approaches to do well on molecular data, and to take more time than distance based methods[56]. For rough analysis, genetic distances as produced by mash can be used for clustering and to produce a rough coarse-grained topology. Consideration of whether to include the accessory genome in the inference or to analyse it separately is important, and will be dependent on the species and lineage being studied.

Choice of method will also depend on why the tree is being built in the first place. If it is for overall population structure, then a more approximate approach will likely suffice, as such analysis is unlikely to delve into precise topology differences at the tips of the tree. All the approaches we recommend were able to recover the

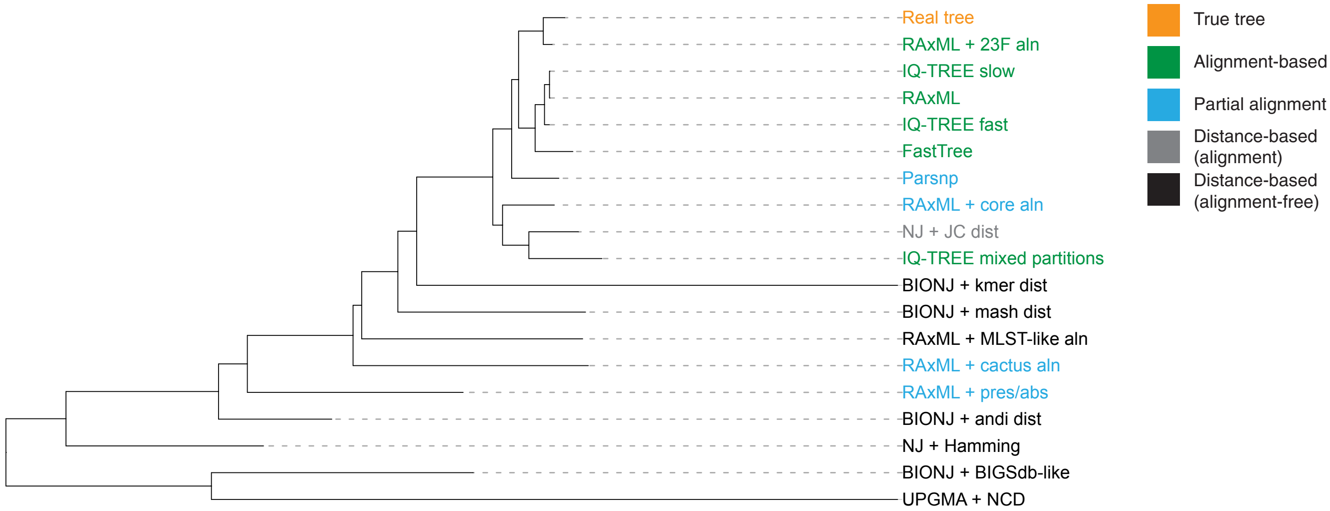

Figure 4. Tree of tree methods. Using the KC metric between all the inferred phylogenies in Table 1 to create a pairwise distance matrix, an NJ tree created from this matrix. This shows how the topologies from all methods are related to each other (a tree-of-trees, or supertree). The true tree is in orange at the top, and four classes of methods are labeled. For alignment-based methods the mapping of reads to the TIGR4 reference was used, unless explicitly stated. We also performed multi-dimensional scaling of these distances in two dimensions to show how the methods clustered (see interactive treespace plots or static Supplementary Figure 6; Supplementary File 1).

correct population clusters with the simulated data. However, for purposes such as transmission cluster inference or association of epidemiological traits (for example a switch in location of isolation) a more precise topology may then be desired.

We also directly compared a range of evolutionary models, run both using BIONJ and ML (Supplementary Table 3; Supplementary File 1). As there are a huge number of sites, and the sites are each low-dimensional, we are much better informed about the site evolution model than the tree. It's easier to get the tree wrong, and hence the inference method used is a more important consideration for tree accuracy. We do note that simpler evolutionary models require less CPU time to run for comparable accuracy. Although maximum likelihood methods cope with missing data much better than distance methods, the extensive missing calls in these simulations (20–40% of sites, due to accessory genes) did not prevent the distance based methods from giving an approximate topology.

For a small number of samples or if computational resources are not a concern, and for phylogenetically focused questions such as model comparison, then a maximum likelihood method is the best choice. However a key point is that in many cases, especially when using a large number of genomes and especially across species with little phylogenetic signal, the phylogeny building software is not the limiting factor in accuracy of the resulting tree. The alignment used is crucial: the quality of sequencing and mapping, whether mobile elements have been masked, and how much confounding signal from recombination and homoplasy can be removed all have important effects on the quality of the final tree. In many cases the observed data are not consistent with a single phylogenetic tree, so rather than aiming for the "best" tree it is important to assess uncertainty in the tree. Bayesian methods are available but are slow and complex[57,58] – we show an example of these on our simulated data in Supplementary Figure 7 (Supplementary File 1). In many cases we would therefore recommend using a faster method such as IQ-TREE's fast mode or FastTree, combined with bootstrap analysis to more efficiently estimate the uncertainty in tree topology[59]. We do note that the bootstrap estimate may be difficult to interpret, as it does not behave as a standard confidence interval due to the implicit assumption that sites are independent[60]. A recent update to the bootstrap may instead be easier to interpret[61], or using the KC metric to compare bootstrap trees[62].

For truly enormous datasets, particularly in cases where producing an alignment is the limiting step, even these approximate methods may prove intractable. In which case using pairwise distances from mash is an alternative approach. One possible problem with mash is that closely related sequences can have a distance of zero, but this can be solved by increasing the sketch size with little extra computational burden. We also note that though the MinHash distance is an approximation, it is a good one, and unlikely to be the limiting factor in these analyses. Instead, accessory genome and mobile elements may be a problem. In these simulations we also tested mash using the core alignment directly, but this resulted in a less accurate tree (KC distance = 71.6); the k-mers sampled by mash do not utilise the information of homology implicit in each column of the alignment.

This work is of course somewhat limited in initial scope. While we tried to choose a true tree with common features, the simulations here are limited, with parameters chosen to model a single species. We also made the choice to ignore branch length differences (though these can as easily be compared) as we think that topological distance is more intuitive, especially for larger differences.

In an age of a bewildering array of options for this analysis and few available direct comparisons we hope that our results are nonetheless instructive, and that these methods can continue to be compared using other benchmark datasets as they appear.

Data availability

Data can be downloaded from the following URLs:

- Code: https://github.com/johnlees/which_tree (GPLv2 license)
- Distances of real gene trees: https://gist.github.com/johnlees/da164a4260e13528e8315e266a46bf3f
- Inferred trees: https://dx.doi.org/10.6084/m9.figshare.5483464[63]
- Interactive treespace plots: https://dx.doi.org/10.6084/m9.figshare.5923300[64]
- Simulation parameters and results (including true alignments of all genes, assemblies and annotations from simulated reads): https://dx.doi.org/10.6084/m9.figshare.5483461[65]

Competing interests

No competing interests were disclosed.

Grant information

This work was supported by the Wellcome Trust (098051). JAL was also supported by a Medical Research Council studentship grant (1365620). CC was supported by the Engineering and Physical Sciences Research Council EPSRC EP/K026003/1 and EPSRC EP/N014529/1.

Supplementary material

Supplementary File 1 - File contain the following supplementary tables and figures:

Click here to access the data.

Supplementary Table 1: Comparison of phylogeny accuracy using all positions versus SNPs plus an ascertainment bias correction for maximum likelihood methods. The KC distance from the true tree, using topology only ($\lambda = 0$) and including branch lengths ($\lambda = 1$) is shown. Resource use, as in Table 1, is shown for each method.

Supplementary Table 2: Twenty gene trees most distant from the core genome tree in 616 *Streptococcus pneumoniae* genomes when using the KC metric with $\lambda = 1$, which only considers branch lengths. The name of the gene, or its name in the *S. pneumoniae* ATCC 700669 genome is shown with the annotated function. Whether each gene was found to be a recombination hotspot in the PMEN1 clone, and whether the hotspot has been specifically described previously are also shown.

Supplementary Table 3: Distance to the true tree for comparable models and methods. Three evolutionary models available both in IQ-tree and SEAVIEW, which were then used to build phylogenies using maximum likelihood (ML) or distances (BIONJ) respectively. Each model has an increasing number of degrees of freedom (df). The KC distances for topology ($\lambda = 0$) and branch length ($\lambda = 1$) are shown, along with the CPU time used for ML inference.

Supplementary Figure 1: An overview of the simulation procedure. Blue boxes show input data: a starting tree and genome at the root, for both evolutionary simulators ALF and DAWG; parameters for each simulator. Orange diamonds show processes: the simulators ALF (for genes) and DAWG (for intergenic regions); perl scripts to combine these results maintaining changes in gene order; pIRS to simulate error-prone reads. Yellow boxes show simulation output data: the full genomes for each sample at the tips of the input tree; aligned sequences for each gene; error-prone reads from the genomes.

Supplementary Figure 2: Applying *plotTreeDiff* between true tree and the closest reconstruction, RAxML + 23F aln (distance 4.35). See top an for explanation of *plotTreeDiff*.

Supplementary Figure 3: Applying *plotTreeDiff* between true tree and one a little further away, the fast IQ-tree (distance 11.3). See top for an explanation of *plotTreeDiff*.

Supplementary Figure 4: Applying *plotTreeDiff* between the true BIGSdb-like (distance 149.8). See top for an explanation of *plotTreeDiff*.

Supplementary Figure 5: Applying *plotTreeDiff* between the true and furthest, UPGMA + NCD (distance 210.5). See top for an explanation of *plotTreeDiff*.

Supplementary Figure 6: A multi-dimensional scaling plot of the distances between all methods projected into two dimensions. This view is zoomed, so the worst methods are outside the plot boundaries.

Supplementary Figure 7: A multi-dimensional scaling plot of the distances between trees sampled from the posterior using mrbayes, projected into two dimensions. There are two chains with different starting points, and the true tree is shown. Both chains appear to have converged on the same regions of treespace (no clustering by colour). There are two favourable modes in this topology space, one of which is closer to the true tree, but less frequently sampled than the other.

References

1. Yang Z: **Computational Molecular Evolution**. OUP Oxford. 2006.
Publisher Full Text

2. Tang P, Gardy JL: **Stopping outbreaks with real-time genomic epidemiology.** *Genome Med.* 2014; **6**(11): 104.
PubMed Abstract | Publisher Full Text | Free Full Text

3. Felsenstein J: **The number of evolutionary trees.** *Syst Biol.* 1978; **27**(1): 27–33.
Publisher Full Text

4. Liu K, Linder CR, Warnow T: **RAxML and FastTree: comparing two methods for large-scale maximum likelihood phylogeny estimation.** *PLoS One.* 2011; **6**(11): e27731.
PubMed Abstract | Publisher Full Text | Free Full Text

5. Zhou X, Shen XX, Hittinger CT, *et al.*: **Evaluating Fast Maximum Likelihood-Based Phylogenetic Programs Using Empirical Phylogenomic Data Sets.** *Mol Biol Evol.* 2018; **35**(2): 486–503.
PubMed Abstract | Publisher Full Text | Free Full Text

6. Bertels F, Silander OK, Pachkov M, *et al.*: **Automated reconstruction of whole-genome phylogenies from short-sequence reads.** *Mol Biol Evol.* 2014; **31**(5): 1077–1088.
PubMed Abstract | Publisher Full Text | Free Full Text

7. Timme RE, Rand H, Shumway M, *et al.*: **Benchmark datasets for phylogenomic pipeline validation, applications for foodborne pathogen surveillance.** *PeerJ.* 2017; **5**: e3893.
PubMed Abstract | Publisher Full Text | Free Full Text

8. Ahrenfeldt J, Skaarup C, Hasman H, *et al.*: **Bacterial whole genome-based phylogeny: construction of a new benchmarking dataset and assessment of some existing methods.** *BMC Genomics.* 2017; **18**(1): 19.
PubMed Abstract | Publisher Full Text | Free Full Text

9. Dalquen DA, Anisimova M, Gonnet GH, *et al.*: **ALF--a simulation framework for genome evolution.** *Mol Biol Evol.* 2012; **29**(4): 1115–1123.
PubMed Abstract | Publisher Full Text | Free Full Text

10. Croucher NJ, Walker D, Romero P, *et al.*: **Role of conjugative elements in the evolution of the multidrug-resistant pandemic clone *Streptococcus pneumoniae*$^{\text{Spain23F}}$ ST81.** *J Bacteriol.* 2009; **191**(5): 1480–1489.
PubMed Abstract | Publisher Full Text | Free Full Text

11. Cartwright RA: **DNA assembly with gaps (Dawg): simulating sequence evolution.** *Bioinformatics.* 2005; **21**(Suppl 3): iii31–38.
PubMed Abstract | Publisher Full Text

12. Kremer PH, Lees JA, Koopmans MM, *et al.*: **Benzalkonium tolerance genes and outcome in *Listeria monocytogenes* meningitis.** *Clin Microbiol Infect.* 2017; **23**(4): 265.e1–265.e7.
PubMed Abstract | Publisher Full Text | Free Full Text

13. Kendall M, Colijn C: **Mapping Phylogenetic Trees to Reveal Distinct Patterns of Evolution.** *Mol Biol Evol.* 2016; **33**(10): 2735–2743.
PubMed Abstract | Publisher Full Text | Free Full Text

14. Cheng L, Connor TR, Sirén J, *et al.*: **Hierarchical and spatially explicit clustering of DNA sequences with BAPS software.** *Mol Biol Evol.* 2013; **30**(5): 1224–1228.
PubMed Abstract | Publisher Full Text | Free Full Text

15. Paten B, Earl D, Nguyen N, *et al.*: **Cactus: Algorithms for genome multiple sequence alignment.** *Genome Res.* 2011; **21**(9): 1512–1528.
PubMed Abstract | Publisher Full Text | Free Full Text

16. Kosiol C, Holmes I, Goldman N: **An empirical codon model for protein sequence evolution.** *Mol Biol Evol.* 2007; **24**(7): 1464–1479.
PubMed Abstract | Publisher Full Text

17. Chen JQ, Wu Y, Yang H, *et al.*: **Variation in the ratio of nucleotide substitution and indel rates across genomes in mammals and bacteria.** *Mol Biol Evol.* 2009; **26**(7): 1523–1531.
PubMed Abstract | Publisher Full Text

18. Chewapreecha C, Harris SR, Croucher NJ, *et al.*: **Dense genomic sampling identifies highways of pneumococcal recombination.** *Nat Genet.* 2014; **46**(3): 305–309.
PubMed Abstract | Publisher Full Text | Free Full Text

19. Hu X, Yuan J, Shi Y, *et al.*: **pIRS: Profile-based Illumina pair-end reads simulator.** *Bioinformatics.* 2012; **28**(11): 1533–1535.
PubMed Abstract | Publisher Full Text

20. Zerbino DR, Birney E: **Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs.** *Genome Res.* 2008; **18**(5): 821–829.
PubMed Abstract | Publisher Full Text | Free Full Text

21. Page AJ, De Silva N, Hunt M, *et al.*: **Robust high-throughput prokaryote *de novo* assembly and improvement pipeline for Illumina data.** *Microb Genom.* 2016; **2**(8): e000083.
PubMed Abstract | Publisher Full Text | Free Full Text

22. Li H: **Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM.** 2013; 3.
Reference Source

23. Li H: **A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data.** *Bioinformatics.* 2011; **27**(21): 2987–2993.
PubMed Abstract | Publisher Full Text | Free Full Text

24. Page AJ, Cummins CA, Hunt M, *et al.*: **Roary: rapid large-scale prokaryote pan genome analysis.** *Bioinformatics.* 2015; **31**(22): 3691–3.
PubMed Abstract | Publisher Full Text | Free Full Text

25. Katoh K, Standley DM: **MAFFT multiple sequence alignment software version 7: improvements in performance and usability.** *Mol Biol Evol.* 2013; **30**(4): 772–780.
PubMed Abstract | Publisher Full Text | Free Full Text

26. Stamatakis A: **RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies.** *Bioinformatics.* 2014; **30**(9): 1312–1313.
PubMed Abstract | Publisher Full Text | Free Full Text

27. Nguyen LT, Schmidt HA, von Haeseler A, *et al.*: **IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies.** *Mol Biol Evol.* 2015; **32**(1): 268–274.
PubMed Abstract | Publisher Full Text | Free Full Text

28. Price MN, Dehal PS, Arkin AP: **FastTree: computing large minimum evolution trees with profiles instead of a distance matrix.** *Mol Biol Evol.* 2009; **26**(7): 1641–1650.
PubMed Abstract | Publisher Full Text | Free Full Text

29. Treangen TJ, Ondov BD, Koren S, *et al.*: **The Harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes.** *Genome Biol.* 2014; **15**(11): 524.
PubMed Abstract | Publisher Full Text | Free Full Text

30. Ondov BD, Treangen TJ, Melsted P, *et al.*: **Mash: fast genome and metagenome distance estimation using MinHash.** *Genome Biol.* 2016; **17**(1): 132.
PubMed Abstract | Publisher Full Text | Free Full Text

31. Haubold B, Klötzl F, Pfaffelhuber P: **andi: fast and accurate estimation of evolutionary distances between closely related genomes.** *Bioinformatics.* 2015; **31**(8): 1169–1175.
PubMed Abstract | Publisher Full Text

32. Lees JA, Vehkala M, Välimäki N, *et al.*: **Sequence element enrichment analysis to determine the genetic basis of bacterial phenotypes.** *Nat Commun.* 2016; **7**: 12797.
PubMed Abstract | Publisher Full Text | Free Full Text

33. Gouy M, Guindon S, Gascuel O: **SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building.** *Mol Biol Evol.* 2010; **27**(2): 221–224.
PubMed Abstract | Publisher Full Text

34. Jolley KA, Maiden MC: **BIGSdb: Scalable analysis of bacterial genome variation at the population level.** *BMC Bioinformatics.* 2010; **11**: 595.
PubMed Abstract | Publisher Full Text | Free Full Text

35. Vitányi PM, Balbach FJ, Cilibrasi RL, *et al.*: **Normalized information distance.** *Information Theory and Statistical Learning.* 2009; 45–82.
Publisher Full Text

36. Alfonseca M, Cebrián M, Ortega A: **Common pitfalls using the normalized compression distance: What to watch out for in a compressor.** *Commun Inf Syst.* 2005; **5**(4): 367–384.
Publisher Full Text

37. Gascuel O: **BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data.** *Mol Biol Evol.* 1997; **14**(7): 685–695.
PubMed Abstract | Publisher Full Text

38. Croucher NJ, Finkelstein JA, Pelton SI, *et al.*: **Population genomics of post-vaccine changes in pneumococcal epidemiology.** *Nat Genet.* 2013; **45**(6): 656–663.
PubMed Abstract | Publisher Full Text | Free Full Text

39. Croucher NJ, Finkelstein JA, Pelton SI, *et al.*: **Population genomic datasets describing the post-vaccine evolutionary epidemiology of *streptococcus pneumoniae*.** *Sci Data.* 2015; **2**: 150058.
PubMed Abstract | Publisher Full Text | Free Full Text

40. Croucher NJ, Campo JJ, Le TQ, *et al.*: **Diverse evolutionary patterns of pneumococcal antigens identified by pangenome-wide immunological screening.** *Proc Natl Acad Sci U S A.* 2017; **114**(3): E357–E366.
PubMed Abstract | Publisher Full Text | Free Full Text

41. Corander J, Fraser C, Gutmann MU, *et al.*: **Frequency-dependent selection in vaccine-associated pneumococcal population dynamics.** *Nat Ecol Evol.* 2017; **1**(12): 1950–1960.
PubMed Abstract | Publisher Full Text | Free Full Text

42. Wernersson R, Pedersen AG: **RevTrans: Multiple alignment of coding DNA from aligned amino acid sequences.** *Nucleic Acids Res.* 2003; **31**(13): 3537–3539.
PubMed Abstract | Publisher Full Text | Free Full Text

43. R Core Team: **R: A Language and Environment for Statistical Computing**. R Foundation for Statistical Computing, Vienna, Austria, 2014.
Reference Source

44. Wickham H: **ggplot2: Elegant Graphics for Data Analysis**. Springer-Verlag New York, 2009.
Publisher Full Text

45. Jombart T, Kendall M, Almagro-Garcia J, *et al.*: **treespace: Statistical exploration of landscapes of phylogenetic trees.** *Mol Ecol Resour.* 2017; **17**(6): 1385–1392.
PubMed Abstract | Publisher Full Text | Free Full Text

46. Lewis PO: **A likelihood approach to estimating phylogeny from discrete morphological character data.** *Syst Biol.* 2001; **50**(6): 913–925.
PubMed Abstract | Publisher Full Text

47. Croucher NJ, Coupland PG, Stevenson AE, *et al.*: **Diversification of bacterial genome content through distinct mechanisms over different timescales.** *Nat Commun.* 2014; **5**: 5471.
PubMed Abstract | Publisher Full Text | Free Full Text

48. Everitt RG, Didelot X, Batty EM, *et al.*: **Mobile elements drive recombination hotspots in the core genome of *staphylococcus aureus*.** *Nat Commun.* 2014; **5**: 3956.
PubMed Abstract | Publisher Full Text | Free Full Text

49. McNally A, Oren Y, Kelly D, *et al.*: **Combined Analysis of Variation in Core, Accessory and Regulatory Genome Regions Provides a Super-Resolution View into the Evolution of Bacterial Populations.** *PLoS Genet.* 2016; **12**(9): e1006280.
PubMed Abstract | Publisher Full Text | Free Full Text

50. Croucher NJ, Harris SR, Fraser C, *et al.*: **Rapid pneumococcal evolution in response to clinical interventions.** *Science.* 2011; **331**(6016): 430–434.
PubMed Abstract | Publisher Full Text | Free Full Text

51. Hedge J, Wilson DJ: **Bacterial phylogenetic reconstruction from whole genomes is robust to recombination but demographic inference is not.** *mBio.* 2014; **5**(6): e02158.
PubMed Abstract | Publisher Full Text | Free Full Text

52. Croucher NJ, Page AJ, Connor TR, *et al.*: **Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using gubbins.** *Nucleic Acids Res.* 2015; **43**(3): e15.
PubMed Abstract | Publisher Full Text | Free Full Text

53. David S, Mentasti M, Tewolde R, *et al.*: **Evaluation of an Optimal Epidemiological Typing Scheme for *Legionella pneumophila* with Whole-Genome Sequence Data Using Validation Guidelines.** *J Clin Microbiol.* 2016; **54**(8): 2135–2148.
PubMed Abstract | Publisher Full Text | Free Full Text

54. Earl D, Nguyen N, Hickey G, *et al.*: **Alignathon: a competitive assessment of whole-genome alignment methods.** *Genome Res.* 2014; **24**(12): 2077–2089.
PubMed Abstract | Publisher Full Text | Free Full Text

55. Zielezinski A, Vinga S, Almeida J, *et al.*: **Alignment-free sequence comparison: benefits, applications, and tools.** *Genome Biol.* 2017; **18**(1): 186.
PubMed Abstract | Publisher Full Text | Free Full Text

56. Guindon S, Gascuel O: **A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood.** *Syst Biol.* 2003; **52**(5): 696–704.
PubMed Abstract | Publisher Full Text

57. Nascimento FF, Reis MD, Yang Z: **A biologist's guide to Bayesian phylogenetic analysis.** *Nat Ecol Evol.* 2017; **1**(10): 1446–1454.
PubMed Abstract | Publisher Full Text | Free Full Text

58. Yang Z, Zhu T: **Bayesian selection of misspecified models is overconfident and may cause spurious posterior probabilities for phylogenetic trees.** *Proc Natl Acad Sci U S A.* 2018; **115**(8): 1854–1859.
PubMed Abstract | Publisher Full Text | Free Full Text

59. Minh BQ, Nguyen MA, von Haeseler A: **Ultrafast approximation for phylogenetic bootstrap.** *Mol Biol Evol.* 2013; **30**(5): 1188–1195.
PubMed Abstract | Publisher Full Text | Free Full Text

60. Efron B, Halloran E, Holmes S: **Bootstrap confidence levels for phylogenetic trees.** *Proc Natl Acad Sci U S A.* 1996; **93**(14): 7085–7090.
PubMed Abstract | Publisher Full Text | Free Full Text

61. Lemoine F, Domelevo Entfellner JB, Wilkinson E, *et al.*: **Renewing Felsenstein's phylogenetic bootstrap in the era of big data.** *Nature.* 2018; **556**(7702): 452–456.
PubMed Abstract | Publisher Full Text

62. Jombart T, Kendall M, Almagro-Garcia J, *et al.*: **treespace: Statistical exploration of landscapes of phylogenetic trees.** *Mol Ecol Resour.* 2017; **17**(6): 1385–1392.
PubMed Abstract | Publisher Full Text | Free Full Text

63. Lees JA: **'which tree' trees.** *Figshare.* 2018.
Data Source

64. Lees JA: **Treespace explorations.** *Figshare.* 2018.
Data Source

65. Lees JA: **Tree simulations.** *Figshare.* 2017.
Data Source

Open Peer Review

Current Referee Status: ✓ ✓ ?

Version 2

Referee Report 30 May 2018

doi:10.21956/wellcomeopenres.15939.r33217

✓

Lauren A. Cowley [1], **Taj Azarian** [2]

[1] Harvard T.H. Chan School of Public Health, Boston, MA, USA

[2] Center for Communicable Disease Dynamics, Department of Epidemiology, Harvard University, Boston, MA, USA

Thank you very much to the authors for addressing all our points and updating the manuscript accordingly, we think the manuscript reads extremely well now.

Competing Interests: The referee Taj Azarian has co-authored a paper with the author Stephen D. Bentley: Azarian, Taj, et al. "Association of Pneumococcal Protein Antigen Serology With Age and Antigenic Profile of Colonizing Isolates." The Journal of infectious diseases 215.5 (2017): 713-722. They do not believe that this has biased their review of the article.

We have read this submission. We believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 1

Referee Report 30 April 2018

doi:10.21956/wellcomeopenres.15526.r32392

?

João A. Carriço

Faculty of Medicine, Institute of Molecular Medicine, University of Lisbon, Lisbon, Portugal

The article by Lees et al. presents us with a very needed evaluation of currently phylogenetic reconstructions methods, based on a simulation based approach. It is a very well written article in a much-needed area and provides several important messages to researchers in this field. I thank the opportunity to review such interesting and important work.

There are however some points that I believe would help the readers in better understanding the details of the analysis, and some further information could help the study reproducibility and replication of results (these are the points that I reported as Partly on my report) and my questions will focus on them.

I will provide the comments per section:

Introduction

Very well written, succinct and full of important and relevant references.

1. (Last sentence of introduction) Concerning providing the code and simulated data, I think all the figshare files and github repositories are in need of a readme file which should contain a better description of the commands and parameters used with some examples to reproduce the paper. Otherwise the claim of reproducibility cannot be made. Even consider a repository for the simulated reads used in this study.

Methods

The methods used by the authors show a tremendous amount of work using several software available to reach their goals. This is highly commendable, but unfortunately also implies that partial description of each software is needed to follow-up without the need of re-reading all the original articles. My following comments are done having this in mind, and to facilitate the reproducibility of the steps:

1. The authors state that they used ALF 1.0 to simulate the evolution along a given phylogenetic tree of 2232 CDS of S. pneumoniae ATCC 700669. I assume that ALF must have some stochastic step and, if such, a seed should be provided to reproduce the same results. Furthermore the phylogeny used was a from a core alignment of Listeria monocytogenes that also has a BAPS classification. At first sight this can be rather confusing for the reader. If I understood correctly, It should be clarified that from the starting CDS, ALF was used to create a final tree with 96 simulated S.pneumoniae strains from the original ATCC 700669, that would correspond to the same topology as the tree from Kremer et al. I also assume that the BAPS groups were recalculated from the final genomes. If so it should be stated on the article.
2. The estimation of the rates to use in GTR the authors used 3 strains (2 pneumo and 1 mitis as outgroup. The claim that this “allowed calculation of SNP and INDEL rates across recent S.pneumoniae evolutionary history” is a bit too extreme and should be moderated.
3. The authors then refer that used DAWG 1.2. to simulate evolution of inter-genic regions. Please clarify how these were defined. The initial text seemed only to refer to the 2232 CDS. Maybe this should be rephrased saying that both CDS and intergenic regions of ATCC 700669 were used in simulating the evolution. Furthermore, the authors should explain how these two approaches can be reconciled in a unique analysis, or at least explicitly state the artificial nature of the result (which I don’t believe that has any impact for the purpose of the paper but should be clarified)
4. Why the choice of velvet for the assembler? Spades has been shown to provide much better results. Furthermore, what were the parameters for velvet? Consider providing the command lines (as supplemental material) for de novo assembly by velvet, for bwa-mem, samtools and roary, as it will be very useful for readers that are new to the field.
5. Consider presenting a summary figure of the whole simulation process, since it would help to guide the reader through the multiple steps done.
6. MLST: why choose 7-genes at random and not use the ones from the schema? I believe that this can have highly misleading results when compared with the defined MLST schema and defining this as MLST analysis mislead the readers.
7. Methods of phylogeny reconstruction and Table 1. Consider numbering the enumeration of methods presented in the text and make a correspondence in a column in Table 1. As it is it is not easy to make the correspondence. For BIGSdb, how was missing data handled and what core schema was used?
8. The Quantification of differences between phylogenetic tree topologies using the KC metric was an excellent choice and the supplemental figures 1 to 3 are really illustrative examples. How are the randomly generated trees generated? This should be added to this section.

9. Core gene trees from real data. The use of MDS to visualize the pair-wise distances is really necessary? An ordered heat-map of the KC metric for the samples would give similar information? I understand the use of the MDS but my feeling is that the final comparison can be biased by the methodology.

Results

1. "We note that all methods except for the NCD were able to recapitulate the population clusters as defined by BAPS" I think this is an important conclusion because in many applications of the trees, researchers compare partitions of the tree and not topology to arrive to their conclusions. In my opinion this should be revisited in the Discussion.
2. Table 1: Add to the table legend the meaning of the "Accessory Genome" column. Also clarify in the text what is the meaning of BAPS 1 and 2. Also explain the meaning of "all" (UPGMA+NCD) in the legend)
3. "However, there is the issue that when a variant is present in a region overlapped by two genes it will be erroneously represented twice." What is a region overlapped by two genes in this context? Were the CDS defined to allow this? This also raises the question what was considered CDS ? Was it what was defined in the previous annotation?
4. Figure 3. See my previous comment to the use of MDS. Also the core genome tree does not behave as average (or centroid) in this dataset and as appears it seems biased to the left of the clusters. I believe that this can be a by-product of the MDS dimensionality reduction. A very interesting result, is what concerns the ribosomal genes. This seems to clearly point out that their use is bad in recapitulating phylogeny and I wonder of this is not only due to the artificial nature of the dataset and similar studies in other species and other might elucidate this matter. It would be interesting to reconcile such results with the results obtained from ribosomal MLST for example in real datasets.
5. "When picking an MLST scheme for an organism, given a choice of genes to use, these phylogenetic signals may be a useful additional consideration." This sentence could be better explained, since it seems really relevant. Could this approach be used as a method to evaluate the choice of MLST target loci for each species?
6. "Although $O(N^2)$ distances need to be evaluated" – You mean N^2 distances. No need for O notation here.
7. On BIGSdb "However, this approach also limits resolution and inference of intra-cluster distances, and produced one of the worst topologies in our tests." Where were the topologies mismatches more common? Within each cluster? Or between clusters? This is relevant because the way information of allelic profiles is commonly used.

Discussion

Well written and informative. The caveats of this study are presented in a paragraph. I think the results of this simulation provide good insights but I wouldn't extrapolate to any other species and dataset. Monomorphic and fastidious species would probably have more similar results using any approach and a study on the impact of mutation and recombination parameters on the final tree-of-trees would be very interesting to see as a future follow-up study.

Is the work clearly and accurately presented and does it cite the current literature?

Partly

Is the study design appropriate and is the work technically sound?

Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Partly

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?
Partly

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Reader Comment 24 May 2018

John Lees, New York University School of Medicine, USA

Approved with Reservations
The article by Lees et al. presents us with a very needed evaluation of currently phylogenetic reconstructions methods, based on a simulation based approach. It is a very well written article in a much-needed area and provides several important messages to researchers in this field. I thank the opportunity to review such interesting and important work.

There are however some points that I believe would help the readers in better understanding the details of the analysis, and some further information could help the study reproducibility and replication of results (these are the points that I reported as Partly on my report) and my questions will focus on them.

Thank you for positive comments, and constructive criticisms below. We respond to each in turn, where the point has not also been raised by one of the other reviewers.

I will provide the comments per section:

Introduction
Very well written, succinct and full of important and relevant references.

(Last sentence of introduction) Concerning providing the code and simulated data, I think all the figshare files and github repositories are in need of a readme file which should contain a better description of the commands and parameters used with some examples to reproduce the paper. Otherwise the claim of reproducibility cannot be made. Even consider a repository for the simulated reads used in this study.

This point has also been made by Dr. Ashton – we have significantly improved this aspect of the data availability. We looked into making simulated reads available, but the size of the data proved too large for the repository. We think that the commands used, as well as the data before and after read generation (including assemblies and mapped alignments) being available will suffice.

Methods
The methods used by the authors show a tremendous amount of work using several software available to reach their goals. This is highly commendable, but unfortunately also implies that partial description of each software is needed to follow-up without the need of re-reading all the original articles. My following comments are done having this in mind, and to facilitate the reproducibility of the steps:

As a general response, we have added the specific commands used to the github repository.

The authors state that they used ALF 1.0 to simulate the evolution along a given phylogenetic tree of 2232 CDS of S. pneumoniae ATCC 700669. I assume that ALF must have some stochastic step and, if such, a seed should be provided to reproduce the same results.

ALF was run with seed = 1. This is in the parameters file, which is now also on the github.

Furthermore the phylogeny used was a from a core alignment of Listeria monocytogenes that also has a BAPS classification. At first sight this can be rather confusing for the reader. If I understood correctly, It should be clarified that from the starting CDS, ALF was used to create a final tree with 96 simulated S.pneumoniae strains from the original ATCC 700669, that would correspond to the same topology as the tree from Kremer et al. I also assume that the BAPS groups were recalculated from the final genomes. If so it should be stated on the article.

We have modified the text to clarify these issues.

The estimation of the rates to use in GTR the authors used 3 strains (2 pneumo and 1 mitis as outgroup. The claim that this "allowed calculation of SNP and INDEL rates across recent S.pneumoniae evolutionary history" is a bit too extreme and should be moderated.

We have removed this statement.

The authors then refer that used DAWG 1.2. to simulate evolution of inter-genic regions. Please clarify how these were defined. The initial text seemed only to refer to the 2232 CDS. Maybe this should be rephrased saying that both CDS and intergenic regions of ATCC 700669 were used in simulating the evolution. Furthermore, the authors should explain how these two approaches can be reconciled in a unique analysis, or at least explicitly state the artificial nature of the result (which I don't believe that has any impact for the purpose of the paper but should be clarified)

We have clarified these issues, and also added more description on the github for the interested reader.

Why the choice of velvet for the assembler? Spades has been shown to provide much better results. Furthermore, what were the parameters for velvet? Consider providing the command lines (as supplemental material) for de novo assembly by velvet, for bwa-mem, samtools and roary, as it will be very useful for readers that are new to the field.

This was also raised in Dr Cowley's review, which we have responded to above. Command lines have been added to the README on github.

Consider presenting a summary figure of the whole simulation process, since it would help to guide the reader through the multiple steps done.

We have added a summary as Supplementary figure 1

MLST: why choose 7-genes at random and not use the ones from the schema? I believe that this can have highly misleading results when compared with the defined MLST schema and defining this as MLST analysis mislead the readers.

We have responded to this along with Dr. Cowley's major comment #3 above, which raises the same issue.

Methods of phylogeny reconstruction and Table 1. Consider numbering the enumeration of methods presented in the text and make a correspondence in a column in Table 1. As it is it is not easy to make the correspondence. For BIGSdb, how was missing data handled and what core schema was used?

Unfortunately the bullet points for the methods in table 1 do not directly correspond to entries in table, as in general there is a combination between an alignment/distance generation and then tree generation method. We think it is clearest to keep table 1 as presented (also in line with figure 4) as we have tried to choose common approaches. We have also kept the bullets in the methods to avoid repetition of e.g. RAxML runs for different input alignments, and make it easier to read than prose. We have added these details for BIGSdb.

The Quantification of differences between phylogenetic tree topologies using the KC metric was an excellent choice and the supplemental figures 1 to 3 are really illustrative examples. How are the randomly generated trees generated? This should be added to this section.

We have added this detail.

Core gene trees from real data. The use of MDS to visualize the pair-wise distances is really necessary? An ordered heat-map of the KC metric for the samples would give similar information? I understand the use of the MDS but my feeling is that the final comparison can be biased by the methodology.

The MDS is useful for an initial exploration of the data, and was useful to prevent biasing by assuming the core genome topology is 'best'. However we take the point that this is the case for this data, and therefore this representation may be favourable. We have therefore added a gist which gives distances of each gene from the core genome tree, and has the advantage of being sortable, searchable and downloadable.

Results

"We note that all methods except for the NCD were able to recapitulate the population clusters as defined by BAPS" I think this is an important conclusion because in many applications of the trees, researchers compare partitions of the tree and not topology to arrive to their conclusions. In my opinion this should be revisited in the Discussion.

See our response to Dr. Azarian's major comment #1 above, which we think covers this issue.

Table 1: Add to the table legend the meaning of the "Accessory Genome" column. Also clarify in the text what is the meaning of BAPS 1 and 2. Also explain the meaning of "all" (UPGMA+NCD) in the legend)

We have added these necessary details.

"However, there is the issue that when a variant is present in a region overlapped by two genes it will be erroneously represented twice." What is a region overlapped by two genes in this context? Were the CDS defined to allow this?

This is a potential issue with genes clustering in general (through roary, cgMLST or wgMSLT), which we wished to point out to readers. For many bacterial genomes, CDS are annotated/defined such that this is possible. Indeed, many genes in bacterial operons overlap, which may be forgotten when looking at individual genes. For tree building it isn't too much of an issue as there are many correlated sites.

This also raises the question what was considered CDS ? Was it what was defined in the previous annotation?

The CDS used for downstream analysis are those found through annotation of assemblies, not the original definition (which may be overly generous, as some regions are hard to consistently assemble and annotate but would appear perfectly in the simulations). We have clarified this in the methods.

Figure 3. See my previous comment to the use of MDS. Also the core genome tree does not behave as average (or centroid) in this dataset and as appears it seems biased to the left of the clusters. I believe that this can be a by-product of the MDS dimensionality reduction. A very interesting result, is what concerns the ribosomal genes. This seems to clearly point out that their use is bad in recapitulating phylogeny and I wonder of this is not only due to the artificial nature of the dataset and similar studies in other species and other might elucidate this matter. It would be interesting to reconcile such results with the results obtained from ribosomal MLST for example in real datasets.

See the comment above re: the use of MDS. For figure 3 the data used is real (not the simulated data). To try and add to this result as suggested we have now made a tree from a concatenation of all these genes (similar to rMLST). A full comparison of these schemes and across species would be great, but beyond the scope of the current paper.

"When picking an MLST scheme for an organism, given a choice of genes to use, these phylogenetic signals may be a useful additional consideration." This sentence could be better explained, since it seems really relevant. Could this approach be used as a method to evaluate the choice of MLST target loci for each species?
We did, in an earlier version of the manuscript, try to use the KC metric to pick 'optimal' MLST schemes. Naturally, as there are roughly 1000C7 =~ 10^{17} such schemes this is a challenging search space, coupled with the fact that for each selection alignments need to be concatenated and trees constructed. We had some success with simulated annealing and genetic algorithms to

solve this problem, with the result that in terms of topology alone there are many schemes which work well, and some genes which should be excluded. This was quite different compared with the main thrust of the paper so we decided to leave it out, but we have added some text here to discuss further.

"Although O(N^2) distances need to be evaluated" – You mean N^2 distances. No need for O notation here.

Thanks, we've fixed this in version two.

On BIGSdb "However, this approach also limits resolution and inference of intra-cluster distances, and produced one of the worst topologies in our tests." Where were the topologies mismatches more common? Within each cluster? Or between clusters? This is relevant because the way information of allelic profiles is commonly used.

We have added supplementary figure 4 to show the differences for this approach specifically.

Discussion
Well written and informative. The caveats of this study are presented in a paragraph. I think the results of this simulation provide good insights but I wouldn't extrapolate to any other species and dataset. Monomorphic and fastidious species would probably have more similar results using any approach and a study on the impact of mutation and recombination parameters on the final tree-of-trees would be very interesting to see as a future follow-up study.

Competing Interests: No competing interests were disclosed.

Referee Report 20 April 2018

doi:10.21956/wellcomeopenres.15526.r32389

✓

Philip M. Ashton
Oxford University Clinical Research Unit, Ho Chi Minh City, Vietnam

The manuscript from Lees and colleagues aims to describe the accuracy and speed of a wide variety of methods for the construction of phylogenetic trees. They achieve this aim in a generally readable and very informative paper. They show that maximum likelihood based approaches using alignment to a closely related reference genome provide the best inference of the simulated true phylogeny. There are various other interesting nuggets spread throughout the paper and it is an interesting read for anyone working with bacterial phylogenies.

I was also interested to note the authors decision to submit to Wellcome Open Research. My hope is that they will take advantage of Wellcome Open Research allowing updating of articles with 'minor' new analyses to include new software which may be released for phylogenetic analysis.

The article is a nice crystallisation and examination of many pieces of received wisdom in bacterial phylogenomics community, especially the balance between accuracy and speed for mash/kmer trees, NJ trees of alignment data and ML trees of alignment data.

I think the work is well presented, well carried out and the conclusions do not over-reach the results.

Minor comments

- In the introduction, this sentence doesn't make sense - 'A more recent, larger study in eukaryotes compared these an IQ-TREE in terms of best likelihood on both species and gene trees'

- As the authors and other reviewers allude to, it is sometimes forgotten that a single tree is not a very realistic representation of the output of an ML phylogenetic analysis. It would be interesting to try and represent this somehow for the different methods. Perhaps a visualisation along the lines of supp figure 4, but with 100 bootstraped trees per method, or the 100 trees with the best ML scores. I appreciate that this is already a busy figure, so I leave it up to the authors whether to do this, or if there is a better way to do it.

- The authors have uploaded scripts to an accompanying github repo, but there is no readme or guide to which scripts relate to which parts of the paper. This should be improved.

Is the work clearly and accurately presented and does it cite the current literature?

Yes

Is the study design appropriate and is the work technically sound?

Yes

Are sufficient details of methods and analysis provided to allow replication by others?

Partly

If applicable, is the statistical analysis and its interpretation appropriate?

Yes

Are all the source data underlying the results available to ensure full reproducibility?

Yes

Are the conclusions drawn adequately supported by the results?

Yes

Competing Interests: No competing interests were disclosed.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reader Comment 24 May 2018

John Lees, New York University School of Medicine, USA

The manuscript from Lees and colleagues aims to describe the accuracy and speed of a wide variety of methods for the construction of phylogenetic trees. They achieve this aim in a generally readable and very informative paper. They show that maximum likelihood based approaches using alignment to a closely related reference genome provide the best inference of the simulated true phylogeny. There are various other interesting nuggets spread throughout the paper and it is an interesting read for anyone working with bacterial phylogenies.

I was also interested to note the authors decision to submit to Wellcome Open Research. My hope is that they will take advantage of Wellcome Open Research allowing updating of articles with 'minor' new analyses to include new software which may be released for phylogenetic analysis.

The article is a nice crystallisation and examination of many pieces of received wisdom in bacterial phylogenomics community, especially the balance between accuracy and speed for mash/kmer trees, NJ trees of alignment data and ML trees of alignment data.

I think the work is well presented, well carried out and the conclusions do not over-reach the results.

Thank you for all the positive comments. We also hope that the availability of our simulated data will allow others to test new methods. We have replied to your suggestions below.

Minor comments

- *In the introduction, this sentence doesn't make sense - 'A more recent, larger study in eukaryotes compared these an IQ-TREE in terms of best likelihood on both species and gene trees'*

We have added the missing words into this sentence.

- *As the authors and other reviewers allude to, it is sometimes forgotten that a single tree is not a very realistic representation of the output of an ML phylogenetic analysis. It would be interesting to try and represent this somehow for the different methods. Perhaps a visualisation along the lines of supp figure 4, but with 100 bootstraped trees per method, or the 100 trees with the best ML scores. I appreciate that this is already a busy figure, so I leave it up to the authors whether to do this, or if there is a better way to do it.*

We think it would be challenging to represent and offer a clear interpretation of the variability of final trees for all the methods used in this manuscript. Instead, we have added trees sampled from the posterior of mrbayes as supplementary figure 7, to try and help illustrate this point further. Further investigation into variation in the posterior, especially with real data, would no doubt be interesting, but beyond the scope of the current work.
We have also added an extra citation to a recent update of the bootstrap and a reference to a paper where this was investigated with bootstrap trees of Dengue virus sequences, for the interested reader. (While indeed true that a single ML tree is not a full representation of a phylogenetic analysis, it is what’s used in almost all publications in this field, perhaps due to the challenge of combining this data either visually or into an informative statistic)

- *The authors have uploaded scripts to an accompanying github repo, but there is no readme or guide to which scripts relate to which parts of the paper. This should be improved.*

Thanks for this point. We have significantly improved the documentation and coverage of the repo, also in line with Dr Carriço’s suggestions below.

Competing Interests: No competing interests were disclosed.

Referee Report 09 April 2018

doi:10.21956/wellcomeopenres.15526.r32391

Lauren A. Cowley [1], **Taj Azarian** [2]

[1] Harvard T.H. Chan School of Public Health, Boston, MA, USA

[2] Center for Communicable Disease Dynamics, Department of Epidemiology, Harvard University, Boston, MA, USA

We have decided to provide a joint review from two postdocs in the Hanage lab of "Evaluation of phylogenetic reconstruction methods using bacterial whole genomes: a simulation based study".

Lauren Cowley's review:
I am grateful for the opportunity to review "Evaluation of phylogenetic reconstruction methods using bacterial whole genomes: a simulation based study" by Lees et al. I found it very interesting and especially useful and relevant to my own research. I am sure I will refer back to it a few times in the future. I also think it will be highly valuable to the wider community of researchers doing bacterial genomic research, which is growing. I think it is likely to become a highly cited paper.

The authors provide a thorough and well thought out analysis of different methods for inferring phylogeny from bacterial datasets. They have also provided details of the computational time and memory required for each method. They have represented their work clearly and produced very informative figures that are extremely useful to the reader. I recommend this paper for publication and just have a few thoughts:

Positive points:

- Very interesting paper
- It's a very useful paper for picking the most appropriate method
- The figures are very nicely produced
- Figure 4 is fantastic.

Major revisions:

1. Would be nice to see this kind of analysis for rooting affects. Could you do alignments with and without an outgroup to show which methods were close to the right rooting/ordering without the outgroup?
2. Another distance matrix software that you could include is called Disty McMatrixface, would be nice to see if there is any variation in that?
3. You state you selected the MLST genes at random, were the genes not checked for being under selective pressure or likely to recombine? MLST genes are not chosen at random and are usually housekeeping genes that are not expected to recombine a lot or be under particular selective pressures. This will affect that aspect of the analysis.
4. It is very intuitive that the genes with discordant trees are recombination hotspots, none of the analysis was run with a post gubbins alignment? Would it not be important to include this?

Minor revisions:

1. You generated error prone illumina reads with pIRS, is there any variation from wgsim?
2. I'm interested that you used velvet instead of SPAdes? I have usually found SPAdes better for bacterial assembly. What assembly parameters did you use? What K size? You say you improved the resulting scaffolds? How? Assembly quality will greatly affect the Parsnp analysis, there should be some mention of that.
3. You state that including the accessory genome is ok in Pneumo but it would be nice to state that for E. coli this is very inadvisable, maybe also give some other examples where the accessory genome would affect this kind of analysis.

4. I like table 1, I think an extra column with recommended use would be helpful. You mention a few times in the text where you might choose that form of analysis but it would be nice to summarise it in the table too.

Taj Azarian's review:

Thank you for the opportunity to review "Evaluation of phylogenetic reconstruction methods using bacterial whole genomes: a simulation based study" by Lees and colleagues. In their analysis, the authors methodically assess a variety of methods to infer a phylogeny from a bacterial genomic dataset. They compare the ability to recover a "true tree" as well of the computational time required using different phylogenetic reconstruction methods. I feel this is much needed work, as we seem to have drifted away from the phylogenetic tree as the central finding, now just an intermediate step in analysis of large bacterial genomic datasets. As such, many often overlook the implication of their alignment and phylogeny inference methods. Overall, the manuscript is consummate. I include some general comments and suggestions below, which I feel would strengthen their manuscript.

1. There should be some focused text in the introduction or discussion about the users goal of phylogenetic reconstruction and how this would possibly determine the analytical approach (i.e., why are you making a tree?). If population structure is your main goal, then almost all approaches will recover the correct level 1 and 2 BAPS clusters. If you are more interested in investigating transmission or the association of epidemiological traits, then perhaps the resolution in tip branch-lengths and topology is important; therefore, an approach that uses a reference-based alignment may be better. It could also be clarified that use of a core-genome alignment at the species level (i.e., not just a lineage/clone) could result in a good amount of signal loss within the BAPS clusters (which is why a lot of these trees have pancaked clades). Further, if more resolution is desired, then reference-based alignment may be performed on a specific BAPS cluster using a close reference (something that is often done in practice). Last, it may be worth including a reminder that violations in some if not all assumptions are made when inferring a phylogeny from a bacterial dataset. As you state, the "true tree" is almost never recovered, but I feel like a lot of researchers forget there are assumptions that are made every time you infer a tree. Certainly, not all of the above needs to be included, but some consideration should be made to incorporating these concepts into the text.

2. It should be stated up front the reason for only comparing tree topology and not branch lengths. I am assuming this was done because branch lengths using distance and character-based tree inference methods would vary, possibly unfairly biasing toward ML trees. In addition, the change in number of sites used would affect branch lengths (core vs reference vs MLST), and none would necessarily be "wrong".

3. Have the authors explored how the true tree topology (regarding the "qualities" they mention) may impact the performance of various phylogeny inference approaches? For example, it is known that UPGMA perform particularly bad in certain situations. The authors mention this in the discussion regarding varying degrees of phylogenetic signal. I would imagine that with low signal, character based methods would perform better than distance-based approaches. Does this matter, or are the errors "washed-out" when using genome-wide data as seen with putative recently admixed genes?

4. Everyone has their own "pet" approach and the authors could spend a lifetime testing different combinations of methods. Having said that, there is one approach that I believe should be evaluated for its possible computational savings. I almost always use RAxML pthreads on SNP alignments using ascertainment bias correction because I have experienced (anecdotally) faster

run times than using the full alignment. My understanding is that using only variant sites will impact branch lengths to some degree (because invariant sites are used in the likelihood calculation) but not the overall tree topology. I think it is worth trying and including if there are significant computational savings to using the full alignment. I would suggest the following: using either the core gene alignment or reference-based alignment, extract variant sites using SNP-sites. Then run RAxML something like this: raxmlHPC-PTHREADS-SSE3 -T 16 -f a -p 12345 -s alignment.fasta -x 12345 -# 100 -m ASC_GTRGAMMA -n alignment --asc-corr=lewis (note-1 this is for v8.2.1 which may be different for 7.8.6. note-2: you can remove the bootstrap option). See if there are memory and CPU time savings and then compare the topology.

Minor comments

- Consider revising the conclusions in the abstract to include the best method for recovering the True Tree (RAxML + reference-based alignment)
- State whether Roary was used with the default PRANK codon aware alignment or mafft alignment. PRANK takes considerably longer (as I am sure you know) and may only perform marginally better in terms of recovering branch lengths.
- In the methods, you mention that ,"Hamming distance between rows of the gene presence/absence matrix produced by Roary (using 95% blast ID cutoff)." Did you use the gene presence/absence output from Roary (accessory_binary_genes.fa only contains a subsample of all accessory COGs) or the entire accessory genome manually extracted from the gene_presence_absence.csv? If the prior, I would consider repeating using the entire presence/absence alignment.
- There are a few sentences that are a little hard to track due to length. For example, in the Methods on page 3, the sentence describing the test tree could be revised as follows: "We identified a phylogeny (Figure 1), originally produced by Kremer et al. from a core genome alignment of 96 *Listeria monocytogenes* genomes from patients with bacterial meningitis, which had a number of qualities we wished to be able to reproduce. Particularly, it possessed two distinct lineages (also making midpoint rooting suitable, and negating the strong dependence on correct rooting implicit in the Kendall and Colijn metric), several clonal groups within each lineage, long branches and a polyphyletic population cluster (population clusters were estimated from a core genome alignment using Bayesian Analysis of Population Structure v6.0 (BAPS)). "
- Another distance approach worth considering would be Torsten Seemens SNP-Dist https://github.com/tseemann/snp-dists since it is rapid and allows for raw SNP distances and simple models (JC, HKY etc).
- The authors state, "a possibility for combining these two data types would be to have separate model partitions for SNP variation and gene gain/loss." This would indeed be very interesting.
- If the authors do consider SNP sites only, I would be interested in how the inclusion of gapped-sites of Ns impacts the results.

Is the work clearly and accurately presented and does it cite the current literature?

Yes

Is the study design appropriate and is the work technically sound?

Yes

Are sufficient details of methods and analysis provided to allow replication by others?

Partly

If applicable, is the statistical analysis and its interpretation appropriate?

Yes

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: The referee Taj Azarian has co-authored a paper with the author Stephen D. Bentley: Azarian, Taj, et al. "Association of Pneumococcal Protein Antigen Serology With Age and Antigenic Profile of Colonizing Isolates." The Journal of infectious diseases 215.5 (2017): 713-722. They do not believe that this has biased their review of the article.

We have read this submission. We believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however we have significant reservations, as outlined above.

Reader Comment 24 May 2018

John Lees, New York University School of Medicine, USA

Lauren Cowley's review:
I am grateful for the opportunity to review "Evaluation of phylogenetic reconstruction methods using bacterial whole genomes: a simulation based study" by Lees et al. I found it very interesting and especially useful and relevant to my own research. I am sure I will refer back to it a few times in the future. I also think it will be highly valuable to the wider community of researchers doing bacterial genomic research, which is growing. I think it is likely to become a highly cited paper.

The authors provide a thorough and well thought out analysis of different methods for inferring phylogeny from bacterial datasets. They have also provided details of the computational time and memory required for each method. They have represented their work clearly and produced very informative figures that are extremely useful to the reader. I recommend this paper for publication and just have a few thoughts:

Positive points:

- *Very interesting paper*
- *It's a very useful paper for picking the most appropriate method*
- *The figures are very nicely produced*
- *Figure 4 is fantastic.*

Thanks for all these kind comments, and also all the constructive points made below. Please find a reply to each point below. Where possible we have incorporated your suggested changes into version two.

Major revisions:

Would be nice to see this kind of analysis for rooting affects. Could you do alignments with and without an outgroup to show which methods were close to the right rooting/ordering without the outgroup?

The KC metric is sensitive to root position – which is one reason why we chose a tree with an obvious/consistent midpoint root from a large separation between two lineages. Indeed, all methods get the same root with this test. When the root is different, the KC metric's distances reflect a different root choice in addition to any differences in the rest of the topology. We note this in the methods '(also making midpoint rooting suitable, and negating the strong dependence on correct rooting implicit in the Kendall and Colijn metric)'.
Another comparison, as suggested, would be with alternate topologies, first checking whether methods are close to getting the same midpoint root where it is less clear where that should be placed. However, as this would involve a very different simulation setup, we think it is probably beyond the scope of the current work.
We have added to the text to address the issue of rooting.

Another distance matrix software that you could include is called Disty McMatrixface, would be nice to see if there is any variation in that?

We have added this to the comparisons, and it is now in table 1 and figure 4.

You state you selected the MLST genes at random, were the genes not checked for being under selective pressure or likely to recombine? MLST genes are not chosen at random and are usually housekeeping genes that are not expected to recombine a lot or be under particular selective pressures. This will affect that aspect of the analysis.

Dr Carriço also mentions this point in his review, this reply is to points made about MLST in both of these reviewer comments.
The annotation of the genes in the simulations does not correspond to their original function and conservation – they do not retain their function. There is therefore no direct analogue of housekeeping genes, and in particular using the genes from the actual MLST scheme is inappropriate as some of them are involved in gene loss events.
We decided the next best thing was to choose a set of seven genes which were conserved and didn't recombine: 'For an MLST alignment we selected seven genes at random from the core alignment (present in all strains) which had not been involved in horizontal transfer events'. The model of evolution we used does not include specific selection or mutation rate by gene (just a gamma distributed rate heterogeneity over all sites in the genome). We think that a random choice from the non-recombining core genes is therefore reasonable as there's not likely to be signal discordant with the phylogeny in some of these genes and not others. Indeed, this choice may even be better than some real MLST schemes which include genes in linkage-disequilibrium with recombining genes (e.g. *ddl* in *S. pneumoniae*) or schemes which included duplicated genes (e.g. *Legionella pneumophila*).
The point we make in the results is reducing in resolution for topology, which to be fair is not what MLST was designed for. For recapitulating population clusters from whole genome sequences a seven-gene scheme works. Nevertheless, we take the point that a direct use of the term 'MLST' may be misleading to the reader, so have changed to 'Seven gene (MLST-like)' or similar throughout.

It is very intuitive that the genes with discordant trees are recombination hotspots, none of the analysis was run with a post gubbins alignment? Would it not be important to include this?

This is a good point, however with the dataset we used we can't use gubbins to remove recombination as the gene alignments are from a species-wide collection. It would certainly be

interesting to explore these tree distances within a lineage where this is possible, but we think this is outside of scope of this paper.
As a side-note we also ran the same simulations with recombination events turned off, and this actually made little difference to the accuracy of the methods (we did not include these results initially to avoid clutter of the comparisons). We have also added a citation which shows the robustness of phylogeny reconstruction to recombination.

Minor revisions:

You generated error prone illumina reads with pIRS, is there any variation from wgsim?

We would expect both methods to be similar, both use similar models for errors in the reads. We chose pIRS as in earlier simulations pIRS worked with the error correction step in SPAdes (doi: 10.1099/mgen.0.000103), whereas this did not always work for other simulation methods.

I'm interested that you used velvet instead of SPAdes? I have usually found SPAdes better for bacterial assembly. What assembly parameters did you use? What K size? You say you improved the resulting scaffolds? How?

We used velvet rather than SPAdes as (compared to SPAdes v3.5) it was more robust – SPAdes can sometimes fail if the coverage histogram cannot be fitted to the data. We have also previously reported a bug that introduced errors (choosing low coverage bases over high coverage bases in bubble resolution) into the assemblies. This was only fixed in version 3.11. We also thought these events might be more likely with simulated data, so opted for a simpler pipeline.

However we still tried to get assemblies of comparable quality to SPAdes by using velvet optimizer (we have now added the link and K size parameters to the methods) and the assembly improvement pipeline in reference 20 (now clearer in the text) with default options. We have previously noted smaller differences between this pipeline and SPAdes than velvet alone (https://doi.org/10.17863/CAM.15617 pages 109-110).

Assembly quality will greatly affect the Parsnp analysis, there should be some mention of that.

We have added in a better description of this issue where parsnp is discussed in the results.

You state that including the accessory genome is ok in Pneumo but it would be nice to state that for E. coli this is very inadvisable, maybe also give some other examples where the accessory genome would affect this kind of analysis.

We had a little discussion of this in this paragraph, but have now referenced specific species examples and added some relevant citations.

I like table 1, I think an extra column with recommended use would be helpful. You mention a few times in the text where you might choose that form of analysis but it would be nice to summarise it in the table too.

Perhaps table 1 is the main thing that readers will see – thanks for the suggestion. We've added a column as suggested.

Taj Azarian's review:

Thank you for the opportunity to review "Evaluation of phylogenetic reconstruction methods using bacterial whole genomes: a simulation based study" by Lees and colleagues. In their analysis, the authors methodically assess a variety of methods to infer a phylogeny from a bacterial genomic dataset. They compare the ability to recover a "true tree" as well of the computational time required using different phylogenetic reconstruction methods. I feel this is much needed work, as we seem to have drifted away from the phylogenetic tree as the central finding, now just an intermediate step in analysis of large bacterial genomic datasets. As such, many often overlook the implication of their alignment and phylogeny inference methods. Overall, the manuscript is consummate. I include some general comments and suggestions below, which I feel would strengthen their manuscript.

Again, thanks for the kind comments and useful suggestions. Please find our replies below.

There should be some focused text in the introduction or discussion about the users goal of phylogenetic reconstruction and how this would possibly determine the analytical approach (i.e., why are you making a tree?). If population structure is your main goal, then almost all approaches will recover the correct level 1 and 2 BAPS clusters. If you are more interested in investigating transmission or the association of epidemiological traits, then perhaps the resolution in tip branch-lengths and topology is important; therefore, an approach that uses a reference-based alignment may be better. It could also be clarified that use of a core-genome alignment at the species level (i.e., not just a lineage/clone) could result in a good amount of signal loss within the BAPS clusters (which is why a lot of these trees have pancaked clades). Further, if more resolution is desired, then reference-based alignment may be performed on a specific BAPS cluster using a close reference (something that is often done in practice). Last, it may be worth including a reminder that violations in some if not all assumptions are made when inferring a phylogeny from a bacterial dataset. As you state, the "true tree" is almost never recovered, but I feel like a lot of researchers forget there are assumptions that are made every time you infer a tree. Certainly, not all of the above needs to be included, but some consideration should be made to incorporating these concepts into the text.

Dr Carriço also makes a similar point in his review. We hoped to make these points, and have now added to the text (in the introduction, results and discussion) to try and emphasise the importance of these issues. These points also contain some of the 'received wisdom' that Dr Ashton references in his review, and are important additions.

It should be stated up front the reason for only comparing tree topology and not branch lengths. I am assuming this was done because branch lengths using distance and character-based tree inference methods would vary, possibly unfairly biasing toward ML trees. In addition, the change in number of sites used would affect branch lengths (core vs reference vs MLST), and none would necessarily be "wrong".

It's true that this was only stated at the end of the discussion, so we have now mentioned this much earlier in the text. The main reason is that distance and ML approaches aren't comparable for branch lengths, and we have mostly considered the question of overall population clusters (as noted in the previous comment) in our discussion.
In some of the extra analysis (within ML approaches e.g. model choice, ascertainment bias) we have compared topology and branch length by also giving the KC metric for lambda = 1.

Have the authors explored how the true tree topology (regarding the "qualities" they mention) may impact the performance of various phylogeny inference approaches? For example, it is known that UPGMA perform particularly bad in certain situations. The authors mention this in the discussion regarding varying degrees of phylogenetic signal. I would imagine that with low signal, character based methods would perform better than distance-based approaches. Does this matter, or are the errors "washed-out" when using genome-wide data as seen with putative recently admixed genes?

These are all good points worth exploring, however comparisons of other possible true tree topologies would require essentially the whole analysis to be repeated multiple times. This is why we tried to pick a starting tree which had some common qualities that are looked for in these kinds of analysis – to try and be as representative as we could with a single starting topology. We don't have the resources to repeat this for different starting trees, so we hope that you will accept this paper as valuable despite this limitation, which we note in the discussion. Perhaps also our improved code availability may also enable the interested researcher to perform their own similar simulations.

Everyone has their own "pet" approach and the authors could spend a lifetime testing different combinations of methods. Having said that, there is one approach that I believe should be evaluated for its possible computational savings. I almost always use RAxML pthreads on SNP alignments using ascertainment bias correction because I have experienced (anecdotally) faster run times than using the full alignment. My understanding is that using only variant sites will impact branch lengths to some degree (because invariant sites are used in the likelihood calculation) but not the overall tree topology. I think it is worth trying and including if there are significant computational savings to using the full alignment. I would suggest the following: using either the core gene alignment or reference-based alignment, extract variant sites using SNP-sites. Then run RAxML something like this: raxmlHPC-PTHREADS-SSE3 -T 16 -f a -p 12345 -s alignment.fasta -x 12345 -# 100 -m ASC_GTRGAMMA -n alignment --asc-corr=lewis (note-1 this is for v8.2.1 which may be different for 7.8.6. note-2: you can remove the bootstrap option). See if there are memory and CPU time savings and then compare the topology.

This is a valuable addition, and is also a question that has arisen before in our groups. We have therefore performed the analysis as suggested, which has been added to the first section of the results and as supplementary table 1. Indeed, we found the using the ascertainment bias and variable sites does give CPU time and memory savings.
This also led us to an inconsistency in the way we ran IQ-TREE and RAxML, which we have now corrected in the text and table 1.

Minor comments

Consider revising the conclusions in the abstract to include the best method for recovering the True Tree (RAxML + reference-based alignment)

We have modified the abstract accordingly.

State whether Roary was used with the default PRANK codon aware alignment or mafft alignment. PRANK takes considerably longer (as I am sure you know) and may only perform marginally better in terms of recovering branch lengths.

We used MAFFT, and have stated this in the methods.

In the methods, you mention that ,"Hamming distance between rows of the gene presence/absence matrix produced by Roary (using 95% blast ID cutoff)." Did you use the gene presence/absence output from Roary (accessory_binary_genes.fa only contains a subsample of all accessory COGs) or the entire accessory genome manually extracted from the gene_presence_absence.csv? If the prior, I would consider repeating using the entire presence/absence alignment.

We actually did not go on use these distances due to a lack of resolution, so have removed this sentence. However in the suggested mixed partition analysis below we used accessory gene presence/absence, though this was extracted from the .Rtab file and therefore contains all COGs.

There are a few sentences that are a little hard to track due to length. For example, in the Methods on page 3, the sentence describing the test tree could be revised as follows: "We identified a phylogeny (Figure 1), originally produced by Kremer et al. from a core genome alignment of 96 Listeria monocytogenes genomes from patients with bacterial meningitis, which had a number of qualities we wished to be able to reproduce. Particularly, it possessed two distinct lineages (also making midpoint rooting suitable, and negating the strong dependence on correct rooting implicit in the Kendall and Colijn metric), several clonal groups within each lineage, long branches and a polyphyletic population cluster (population clusters were estimated from a core genome alignment using Bayesian Analysis of Population Structure v6.0 (BAPS)). "

We have split up the many clauses of this confusing sentence.

Another distance approach worth considering would be Torsten Seemens SNP-Dist https://github.com/tseemann/snp-dists since it is rapid and allows for raw SNP distances and simple models (JC, HKY etc).

This package seems to be one of four written to rapidly produce a Hamming distance/ANI from SNP alignments (one being Disty McMatrixface mentioned by Dr Cowley above, the others being panito by Andrew Page and pairwise_snp_distances by Anders Goncalves de Silva). We've added in one of these, Disty McMatrixface, as a comparison as noted above (selection of this method from the four possibilities was based on its name).

The authors state, "a possibility for combining these two data types would be to have separate model partitions for SNP variation and gene gain/loss." This would indeed be very interesting.

We have added this analysis in (using IQ-tree), though found it to be less accurate. One possible issue with this for real data is including genes that are discordant with the phylogenetic signal from vertical evolution (e.g. mobile genetic elements). Another reason may be roary incorrectly clustering (both false positive/negative clusterings). We have noted this in the updated text.

If the authors do consider SNP sites only, I would be interested in how the inclusion of gapped-sites of Ns impacts the results.

We included gaps in the analysis throughout (which were numerous, in regions where the simulated sequences were distant from the reference being mapped to).

Competing Interests: No competing interests were disclosed.

Discuss this Article

Version 1

Author Response 24 May 2018

John Lees, New York University School of Medicine, USA

We would like to thank all four reviewers for constructive, polite and rapid responses. Their comments have been very helpful in putting together an improved revision. We have added any new trees or interactive plots to the figshare links cited.

One general point: we have tried to keep the manuscript mostly as a straight-up comparison of the methods to make it easier for analysts to see the take-home points of table 1. All the reviewers have made useful suggestions about ways in which the analysis could be extended and ways in which some points could be further explored. For some of these, even if we have made the suggested comparison/extension, we have opted to keep detailed discussion of these points out of the main text to avoid over-complicating the main conclusions. We hope that the comment format of Wellcome Open Research will help readers interested in this more in-depth discussion explore these issues with us and the reviewers.

Competing Interests: No competing interests were disclosed.