Exploring SNP Filtering Strategies: The Influence of Strict vs Soft Core

Mona L. Taouk[1*], Leo A. Featherstone[2], George Taiaroa[1], Danielle J. Ingle[2], Timothy P. Stinear[2], Ryan R. Wick[2*]

1. *Department of Infectious Diseases, The University of Melbourne at the Peter Doherty Institute for Infection and Immunity, Melbourne, Victoria, Australia*
2. *Department of Microbiology and Immunology, The University of Melbourne at the Peter Doherty Institute for Infection and Immunity, Melbourne, Victoria, Australia*

**** Corresponding authors***
Ms Mona Taouk: taouk@student.unimelb.edu.au
Dr Ryan Wick: ryan.wick@unimelb.edu.au

Keywords
phylogenetic; temporal signal, core alignments; SNP thresholds

ABSTRACT

Phylogenetic analyses are crucial for understanding microbial evolution and infectious disease transmission. Bacterial phylogenies are often inferred from single nucleotide polymorphism (SNP) alignments, with SNPs as the fundamental signal within these data. SNP alignments can be reduced to a 'strict core' containing only sites present in all samples. However, as sample size and genome diversity increase, a strict core can shrink markedly, resulting in lost informative data. Here, we propose and provide evidence to support the use of a 'soft core' that tolerates some missing data, preserving more information for phylogenetic analysis. Using large datasets of *Neisseria gonorrhoeae* and *Salmonella enterica* serovar Typhi, we assess different core thresholds. Our results show that strict cores can drastically reduce informative sites compared to soft cores. In a 10,000-genome alignment of *S.* Typhi, a 95% soft core yielded 10 times more informative sites than a 100% strict core. Similar patterns were observed in *N. gonorrhoeae*. We further evaluated the accuracy of phylogenies built from strict- and soft-core alignments using datasets with strong temporal signals. Soft-core alignments generally outperformed strict cores in producing trees that support clock-like behaviour; for instance, the *N. gonorrhoeae* 95% soft-core phylogeny had a root-to-tip regression R^2 of 0.50 compared to 0.21 for the strict-core phylogeny. This study suggests that soft-core strategies are preferable for large, diverse microbial datasets. To facilitate this, we developed *Core-SNP-filter* (github.com/rrwick/Core-SNP-filter), a tool for generating soft-core alignments based on user-defined thresholds.

IMPACT STATEMENT

This study addresses a major limitation in modern microbial genomics – the significant data loss observed in large datasets for phylogenetic analyses, often due to strict-core SNP alignment approaches. As microbial genome sequence datasets grow and variation increases, a strict-core approach can greatly reduce the number of informative sites, compromising phylogenetic resolution. Our research highlights the advantages of soft-core alignment methods which tolerate some missing data and retain more genetic information. To streamline the processing of alignments, we developed *Core-SNP-filter* (github.com/rrwick/Core-SNP-filter), a publicly available resource-efficient tool that provides informative and core sites.

DATA SUMMARY

No new reads were generated for this study. All reads used in this study are publicly available and accessions can be found in Supplementary Dataset 1. Supplementary methods and all code can be found in the accompanying GitHub repository: (github.com/mtaouk/Core-SNP-filter-methods).

INTRODUCTION

Appropriately resolving the phylogenetic relationships between organisms is a fundamental aspect of many scientific disciplines, including ecology, microbiology and evolutionary biology (1). Defining phylogenetic relationships using whole-genome sequencing (WGS) serves as a cornerstone application in clinical and public health settings for tracking the spread of pathogens and pinpointing outbreaks (2-7).

Conventionally, phylogenetic studies using WGS rely on reference-based techniques. In this approach, single nucleotide polymorphisms (SNPs) are identified for each sample by aligning sequencing reads against a reference genome and calling variants using bioinformatic tools such as *Snippy* (github.com/tseemann/snippy). These results are then combined to generate a whole-genome pseudoalignment ('alignment' for brevity) encompassing all relevant samples. If appropriate, recombination events and problematic areas of the genome can be masked from the alignment using tools such as Gubbins or ClonalFrameML (8, 9). This alignment is then traditionally refined to produce what is known as a 'strict core', which includes only the SNP sites present in 100% of samples (see github.com/tseemann/snippy as an example). This strict core approach is commonly used to generate the data required for phylogenetic inference, including SNP distances to infer the relatedness of genomes and patterns of transmission, with these alignments also being the foundation for temporal phylogenetic inferences, especially in outbreak investigations and evolutionary studies (10-17).

The rapid decrease in sequencing costs has revolutionised our capacity for large-scale whole-genome analyses (18). Contemporary microbial studies now routinely involve

WGS for thousands of samples, and public datasets number in the hundreds of thousands for some species. This necessitates revisiting methods for large-scale analyses, both within and across species. This is evidenced by studies employing methods other than whole-genome alignments to determine the genetic relatedness of isolates (19-22), with questions arising regarding the informativeness of alignments and the challenges of dealing with increasingly large datasets (23). As more genomes are included, the strict-core alignment, consisting of sites present in all genomes, can only diminish in size due to the increasing likelihood of missing data in at least one genome, thereby reducing the overall resolution. Some studies have addressed this by employing a 'soft-core' alignment, able to include SNP sites present in fewer than all samples (<100%), although these approaches are not routine (6, 24, 25).

In this study, we evaluate the effectiveness of strict- and soft-core approaches for generating alignments and phylogenetic inferences using four priority bacterial species. To do so, we introduce *Core-SNP-filter* (github.com/rrwick/Core-SNP-filter), a tailored tool which allows for efficient manipulation of large SNP alignments. Through these analyses, we aim to illuminate best practices for researchers seeking to accurately resolve phylogenetic relationships in bacteria.

METHODS

Development and implementation of Core-SNP-filter

To allow for the efficient filtering of core SNPs to user-defined thresholds, we developed *Core-SNP-filter*, which takes a FASTA alignment as input and outputs a reduced FASTA alignment based on core filtering settings. *Core-SNP-filter* can discard invariant columns (controlled by the -e option) and non-core columns (controlled by the -c option). A column is determined to be invariant if the number of unique unambiguous bases (A, C, G or T) in that column is zero or one. A column is determined as core if the fraction of bases in that column which are unambiguous exceeds the user-supplied value (e.g. -c 0.95). All characters other than the unambiguous bases (e.g. N, -, X, etc.) are treated the same.

Core-SNP-filter is implemented in Rust and is resource-efficient. It is available on GitHub (github.com/rrwick/Core-SNP-filter) with thorough usage instructions and a sample dataset.

Large-scale genome analysis using PathogenWatch data

For two priority pathogens, *N. gonorrhoeae* (n = 11,129) and *Salmonella enterica* serovar Typhi (n = 10,991), we downloaded the reads for all genomes available on PathogenWatch, an online database for processing and visualising microbial genome sequences in phylogenetic and geographical contexts with associated accessions and published studies (details in Supplementary Dataset) (26). *Snippy* v4.6.0 was used with default settings to call variants, using *N. gonorrhoeae* NCCP11945 (accession:

CP001050.1) and *S.* Typhi str. CT18 (accession: NC_003198.1) as reference genomes (27, 28).

For both species, *Snippy-core* was used to generate an alignment of 10,000 random genomes. To simulate datasets with varying numbers of genomes, we created subsets of the alignment in increments of 250, ranging from 10,000 to 250 genomes, and in increments of 25, ranging from 100 to 25 genomes, resulting in a total of 44 alignments with different genome counts. For each of these 44 alignments, a core alignment was produced at thresholds of 50%, 60%, 70%, 80%, 90%, 95%, 96%, 97%, 98%, 99% and 100% using *Core-SNP-filter*, once removing invariant sites and a second time keeping all invariant sites. From each, the alignment length (number of remaining sites) was extracted. This entire process was repeated ten times for each species, with a new random seed used for each repetition, and the reported results are the mean of all ten replicates.

To investigate the effect of changing between strict- and soft-core SNP thresholds on genomic epidemiology studies for *N. gonorrhoeae*, core-SNP alignments were generated for 22 individual studies at both 100% and 95% thresholds (7, 12, 17, 29-47). Genomic reads were downloaded from PathogenWatch corresponding to the published studies. For each study, variants were called with *Snippy* using the *N. gonorrhoeae* NCCP11945 reference genome, a pseudoalignment was made with *Snippy-core*, and *Core-SNP-filter* was used to make 100% strict-core and 95% soft-core alignments. The number of variant sites in each core alignment was then counted and compared between the two thresholds.

Evaluation of core-SNP thresholds for phylodynamics with public datasets

Four publicly available datasets were used to test core-SNP thresholds on a temporal regression analysis: 154 *Streptococcus pneumoniae* genomes (48), 192 *N. gonorrhoeae* genomes (12), 88 *Staphylococcus aureus* genomes (10) and 96 *Salmonella enterica* serovar Kentucky genomes (14). These datasets were chosen because they each span a range of dates and contain a strong temporal signal, as demonstrated in their original publications.

For each dataset, we downloaded all available reads and the reference genome used in the dataset's publication (see Supplementary Methods). Quality checks were performed on the reads from the *S. pneumoniae* dataset (excluding samples unable to assemble well, see Supplementary Methods), as these reads were from 2011 and lower quality than the other three datasets. *Snippy* v4.6.0 was used with default settings to produce whole-genome pseudoalignments using the reference genome. *Snippy-core* was used to merge these into a combined alignment for each study. *Gubbins* v3.2.1 was used to generate a pseudoalignment with putative recombinant regions masked (8). We then used *Core-SNP-filter* to produce an invariant-free alignment at each possible

core-SNP threshold (e.g. 0/n, 1/n, 2/n, ... (n-1)/n, n/n) for both the pre-Gubbins and post-Gubbins alignment.

We then used *IQ-TREE* v2.0.3 to produce three independent trees from each pseudoalignment using the TVM+F+ASC+R2 substitution model, the best-supported model identified by *IQ-TREE's* inbuilt ModelFinder throughout these datasets (49). Each tree was rooted by selecting the root position maximising R^2 for root-to-tip regression using *ape* v5.7.1 (50). For each pseudoalignment/tree, we extracted the number of variant sites and R^2 of the root-to-tip regression (a measure of the temporal signal strength for a strict molecular clock).

Benchmarking of Core-SNP-filter against similar tools

To evaluate the performance of *Core-SNP-filter*, we conducted benchmarking tests relative to three other tools: *Goalign* v0.3.7, *SNP-sites* v2.5.1 and *TrimAl* v1.4.1 (51-53). We quantified the time and RAM required for two tasks: converting full alignments to no-invariant-sites alignments and converting no-invariant-sites alignments to 95% core alignments. *Core-SNP-filter* and *Goalign* can perform both tasks, while SNP-sites can only perform the first task and *TrimAl* can only perform the second. Each tool was cloned from GitHub and built from source on a system equipped with an AMD EPYC 7742 CPU and 512 GB of RAM. The test dataset consisted of a 10,000-genome *S.* Typhi pseudoalignment, which was 4,809,037 base pairs in length and 47 GB in size. The sequences contained the characters -, A, C, G, T, N, a, c, g, t and n. Genome counts ranging from 10 to 10,000 were tested, spaced evenly on a logarithmic scale.

The benchmarking process involved subsampling the initial alignment to create datasets of varying genome counts. For each dataset, we used the tools to remove invariant sites and then to produce a 95% core alignment. Each test was repeated eight times, and the minimum time and RAM usage were recorded. MD5 checksums of the output files ensured consistency of results across different tools.

Statistical analyses and data visualisation

All statistical analyses were performed in *RStudio* (v2022.12.0+353) using *R* (v4.2.2). All figures were made using *ggplot2* (v3.4.0).

RESULTS

Revisiting the appropriateness of strict core filtering for large and diverse datasets

We evaluated the impact of using a 100% strict core on alignment size and number of informative sites for phylogenetic analyses for two species with more than ten thousand publicly available genomes: *N. gonorrhoeae* and *S.* Typhi. For *S.* Typhi, we analysed 10,991 genomes collected between 1935 and 2020 from 95 countries, forming a geographically and temporally diverse dataset (accessions in Supplementary Dataset

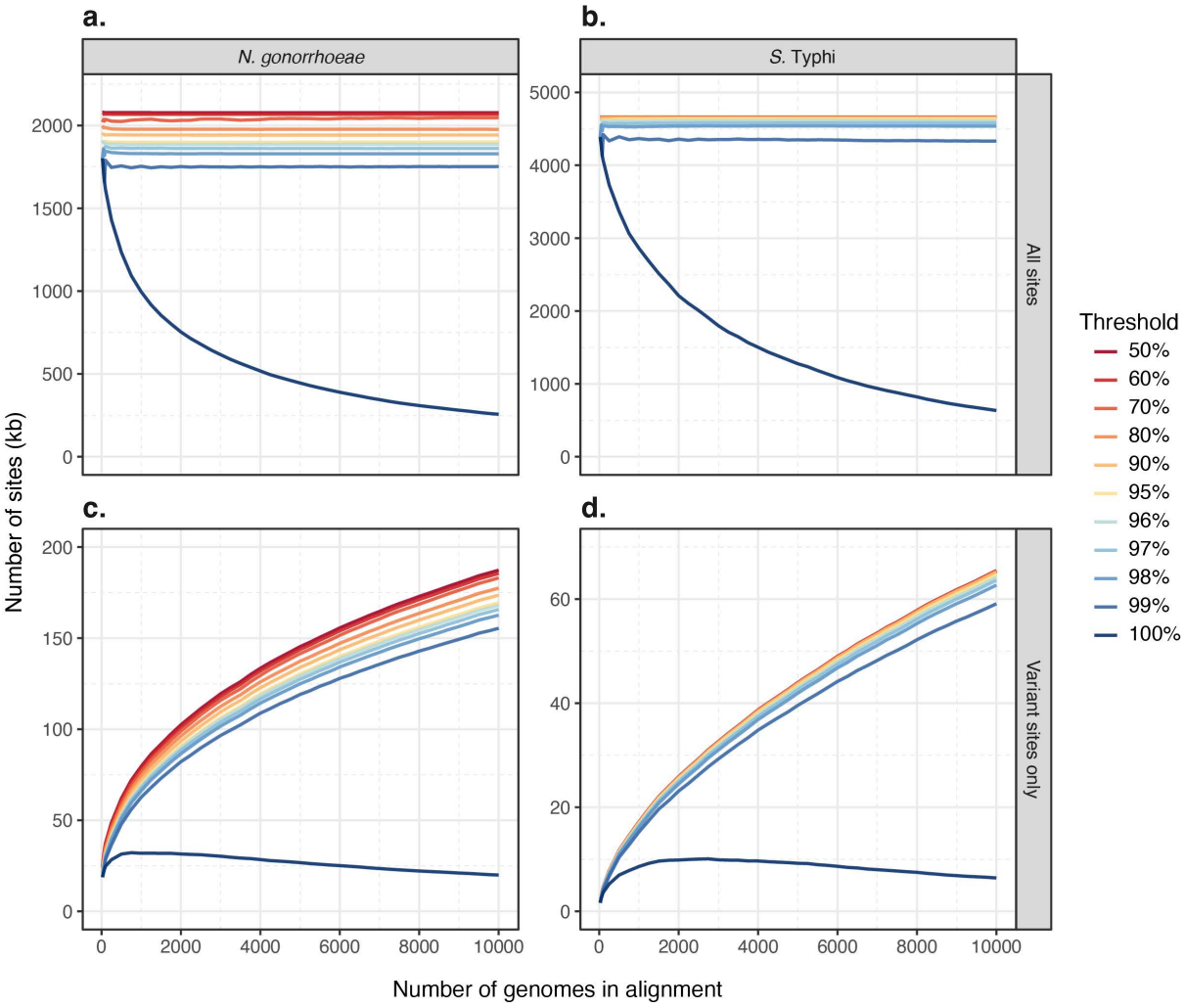
1). These genomes were sequentially subset into 44 alignments varying by number of genomes (see Methods for details). Similarly, for *N. gonorrhoeae*, 11,129 genomes collected between 1928 and 2018 from 62 countries were used to generate ten random alignments of 10,000 genomes each, which were also subset into 44 alignments varying by number of genomes (accessions in Supplementary Dataset 1).

For both *N. gonorrhoeae* and *S.* Typhi, the 100% strict core resulted in progressively fewer alignment sites as the number of genomes increased (Fig. 1a-b, Supplementary Dataset 2). At all other core thresholds tested (50–99%), the number of sites remained stable as the number of genomes increased. For example, an alignment of 10,000 genomes at a strict threshold of 100% resulted in a mean of 256,124 and 633,715 sites for *N. gonorrhoeae* and *S.* Typhi, respectively, compared to a soft threshold of 95%, which resulted in a mean of 1,900,987 and 4,634,869 sites in *N. gonorrhoeae* and *S.* Typhi, respectively.

A similar trend was observed when considering variant sites. For both *N. gonorrhoeae* and *S.* Typhi, the 100% strict core resulted in drastically fewer variant sites compared to other thresholds at all alignment sizes (Fig 1c-d). Except for the 100% strict core, the number of variant sites increased with the number of genomes at all tested thresholds, and the number of variant sites increased as the core threshold decreased. For example, an alignment of 10,000 genomes at a strict threshold of 100% resulted in a mean of 19,926 and 6,415 variant sites output for *N. gonorrhoeae* and *S.* Typhi, respectively, compared to a soft threshold of 95%, which resulted in a mean of 169,409 and 64,711 variant sites output in *N. gonorrhoeae* and *S.* Typhi, respectively. These findings indicate that considerably more evolutionary signal can be captured using soft-core approaches (<100%) compared to strict-core approaches (100%).

In genomic epidemiology, datasets from individual studies may exhibit reduced diversity and potential biases due to variable temporal, contextual and/or geographic sampling. To assess the potential impacts of varying core thresholds, we collated a set of *N. gonorrhoeae* studies (n=22) and generated alignments for each using both 100% and 95% thresholds (see Methods) (7, 12, 17, 29-47). For all but one study, the number of variant sites output increased in the 95% threshold alignment compared to the 100% threshold alignment (Supplementary Fig 1, Supplementary Table 1). Williamson et al. 2019 (2,162 genomes) had the highest increase in variant sites at 511.2%, indicating that large and diverse datasets may particularly benefit from lowering the core threshold (7). Other studies with substantial increases in the number of variant sites included Thomas et al. 2019 (639 genomes, 107.16%) and Alfsnes et al. 2020 (806 genomes, 122.76%) (17, 29). Fifer et al. 2018 (97 genomes) and Kwong et al. 2018 (75 genomes) exhibited smaller increases of 8.59% and 6.16%, respectively, showing that smaller and/or less diverse datasets may see less benefit from lowering the threshold (36, 38). Ezewudo et al. 2015 (18 genomes) showed no change in SNP sites, highlighting that for very small datasets lacking in diversity, threshold adjustments may not impact variant site counts (35).

Figure 1: Evaluation of *Core-SNP-filter* Thresholds on Alignment Size and Variant Sites



The number of alignment sites (in kilobases) is plotted against the number of genomes included in each alignment (ranging from 25 to 10,000). Each alignment was processed at core thresholds ranging from 50% to 100% (indicated by line colour). (a) All sites in N. gonorrhoeae alignments, (b) all sites in S. Typhi alignments, (c) variant sites only in N. gonorrhoeae alignments, and (d) variant sites only in S. Typhi alignments. The number of sites was similar across different core thresholds, except for the 100% threshold, which resulted in far fewer sites, particularly at higher genome counts.

Appropriateness of using a strict core threshold for small datasets with strong temporal signal

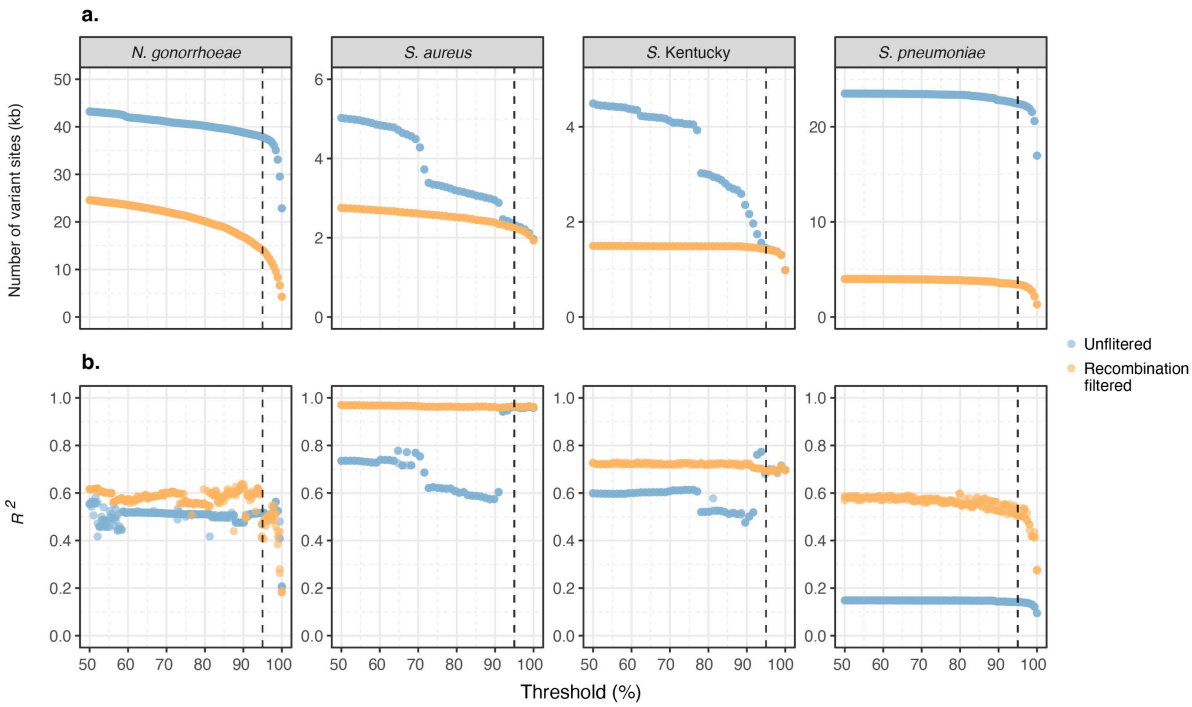
To investigate the effect of varying core thresholds in situations where removing certain sites might be beneficial—such as those inconsistent with the clonal frame or regions affected by recombination—we analysed four existing datasets. These datasets, characterised by a strong temporal signal and limited geographic diversity, provide an ideal context for assessing the appropriateness of different core thresholds for uncovering stronger temporal signal, which is essential for subsequent phylodynamic analyses aiming to make epidemiological inference from time-stamped genome sequences (54, 55). For this analysis, we selected four datasets representing different pathogens: *S. pneumoniae, N. gonorrhoeae, S.* Kentucky and *S. aureus* (10, 12, 14, 48). For each, we generated both pre-Gubbins (unfiltered) and post-Gubbins (recombination-filtered) alignments, which were processed with core thresholds ranging from 50% to 100% (see Methods).

In general, recombination-free alignments had fewer variant sites at each core threshold across all four datasets (Fig. 2a). Similar to the results observed in the large datasets, the number of variant sites for all recombination-free alignments were generally consistent until a decrease as the core threshold approached 100% (Supplementary Dataset 3). The unfiltered alignments had an overall higher number of variant sites, especially at lower core thresholds. Across all datasets, for both the recombination-filtered and unfiltered alignments, the lowest number of variant sites were observed in the 100% core threshold alignments. Additionally, when comparing the tree topology generated from a 100% core-threshold alignment to those from a 95% core-threshold alignment, we observed more polytomous branching in the 100% core alignment (Supplementary Fig. 2). This suggests that the evolutionary relationships between genomes were less resolved at the 100% core-threshold.

Here, R^2 is a key measurement of temporal signal in each dataset, which is essential for subsequent phylodynamic analyses aiming to estimate rates of epidemiological transmission. For the *N. gonorrhoeae* dataset, the R^2 values remained relatively stable across all tested core thresholds, until a steep decline as the core threshold approached 100%, with the same trend being observed for both the recombination-filtered and unfiltered phylogenies (Figure 2b, Supplementary Dataset 3). Similarly, the *S. pneumoniae* R^2 values remained relatively stable across core thresholds, until a steep decline as the core threshold approached 100%. For the *S. pneumoniae* dataset, the impact of removing recombination on the R^2 values was more pronounced, with a mean R^2 difference of 0.41 between the recombination-free and regular phylogenies. In the case of the *S. aureus* dataset, there was a noticeable difference between the R^2 values for the recombination-filtered and unfiltered phylogenies. Across all core thresholds, the recombination-filtered *S. aureus* phylogenies had very consistent R^2 values, while the R^2 values for the unfiltered *S. aureus* phylogenies were much more varied and did not follow a discernible trend. However, at the 100% core threshold, both the recombination-filtered and unfiltered *S. aureus* phylogenies had an R^2 value

of 0.95. Similarly, the *S.* Kentucky dataset also showed relatively consistent R^2 values across all tested core thresholds for the recombination-filtered phylogenies and more varied R^2 values for the unfiltered phylogenies.

Figure 2: Impact of Core-SNP Thresholds on Phylogenetic and Temporal Signal Across Bacterial Datasets



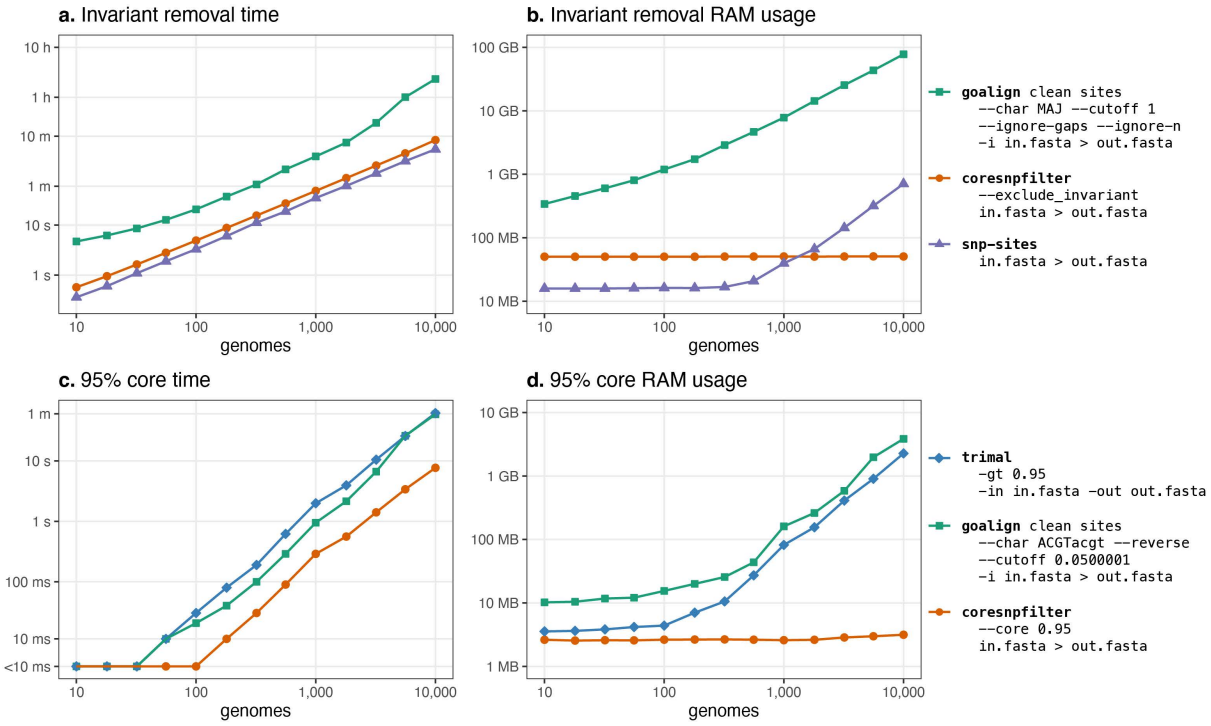
(**a**) The number of variant sites (in kilobases) at each core genome threshold from 50% to 100% for four bacterial datasets: *N. gonorrhoeae*, *S. aureus*, *S.* Kentucky, and *S. pneumoniae*. The variant sites are shown for both recombination-filtered (using Gubbins) and unfiltered alignments. (**b**) The R^2 value from root-to-tip regression at each core threshold from 50% to 100% for phylogenetic trees constructed from the same four bacterial datasets, indicating the strength of the temporal signal. For all plots, the dotted line represents the 95% core threshold. Each dataset shows a steep decline in variant sites as the threshold approaches 100%, with *N. gonorrhoeae* and *S. pneumoniae* also exhibiting a reduction in R^2 as the threshold approaches 100%.

Benchmarking Core-SNP-filter against existing tools

To evaluate the performance of *Core-SNP-filter* against three tools that perform related functions, *Goalign*, *SNP-sites* and *TrimAl*, we conducted benchmarking tests on two specific tasks: converting full alignments to no-invariant-sites alignments and converting no-invariant-sites alignments to 95% core alignments. *Core-SNP-filter* and *Goalign* can perform both tasks, whereas *SNP-sites* can only perform the first task, and *TrimAl* can only perform the second. These tools were chosen because they are written in fast, compiled programming languages: *Core-SNP-filter* in Rust, *Goalign* in Go, *SNP-sites* in C and *TrimAl* in C++.

There was a clear relationship between genome count and execution time for all tools and tasks (Fig. 3). For the task of removing invariant sites, *SNP-sites* was the fastest, with *Core-SNP-filter* taking about 1.5 times longer and *Goalign* taking about seven times longer (Fig. 3a). For the task of producing a 95% core alignment, *Core-SNP-filter* was the fastest, while *Goalign* and *TrimAl* took approximately four times longer (Fig 3c). In terms of RAM usage, *Core-SNP-filter* consistently used a fixed amount of RAM regardless of genome count for both removal of invariant sites (mean: 51.8 MB) and generating the 95% core alignments (mean: 2.8 MB) (Fig 3b and 3d). For removal of invariant sites, both *Goalign's* and *SNP-sites'* RAM usage increased with genome count, reaching a mean of 78.8 GB and 709.1 MB, respectively, for the largest datasets (10,000 genomes). Similarly, for both *Goalign* and *TrimAl*, generating a 95% core threshold alignment resulted in increased RAM usage proportional to genome count, reaching a mean of 4.2 GB and 2.3 GB, respectively, for the largest datasets.

Figure 3: Benchmarking of *Core-SNP-filter* Against Other Tools with Similar Functionality



The time taken (**a**) and RAM usage (**b**) to remove invariant sites from alignments with varying numbers of genomes using three different tools (*Goalign*, *Core-SNP-filter* and *SNP-sites*). The time taken (**c**) and RAM usage (**d**) to generate a 95% core alignment across alignments with varying numbers of genomes using three different tools (*TrimAl*, *Goalign and Core-SNP-filter*). All axes are transformed with a log scale.

DISCUSSION

This study highlights the importance of considering soft-core strategies as a best practice for phylogenetic analyses of microbial datasets. Our findings reveal the limitations of using a strict (100%) core threshold, particularly for large and diverse datasets. Even a slight relaxation from a 100% core to 99% notably increases the number of output sites. A 95% core emerges as a practical choice for many scenarios, striking a balance between retaining informative phylogenetic signals and excluding data inconsistent with the clonal frame of evolution or influenced by recombination. We also benchmarked the performance of our tool, *Core-SNP-filter*, against available tools that perform part, or all, of the required processing, demonstrating its efficiency in handling large datasets.

Traditional methods often rely on generating strict alignments from all genomes in a dataset as the basis for phylogenetic inferences and genomics evolutionary studies. However, as we demonstrate, this approach becomes increasingly impractical and

ineffective as datasets grow. The strict-core method, which requires alignment of SNPs present in all genomes, often results in significant data loss, reducing the resolution and accuracy of phylogenetic analyses. Recent studies have explored alternative methods for generating phylogenies from large microbial datasets, including concatenated core genome multilocus sequence typing (cgMLST) alignments and neighbour-joining trees based on pairwise core SNP distances (19, 56). While these methods address some challenges, our study presents a simple tool striking a balance between traditional strict-core approaches and addressing present limitations. These soft-core approaches are being implemented in a number of large-scale phylogenetic studies (6, 24, 25).

In public health and hospital outbreak settings, phylogenetic analysis is crucial for inferring transmission dynamics. Typically, SNP distances calculated from strict-core alignments are used to determine transmission clusters. However, this method can be problematic due to its sensitivity to reference genome selection, masking of recombinogenic or prophage regions, and the approach taken (23). These issues are exacerbated in large datasets, where a static SNP distance threshold for transmission clustering may be inappropriate. Such thresholds, while standard in many transmission studies and accredited pipelines, fail to account for dynamic core sizes that result from strict-core approaches, particularly in epidemiological studies involving expanding datasets (57-59). Our findings indicate that as datasets expand, the alignment size diminishes under a strict 100% core threshold, potentially leading to over-supported transmission networks and misleading clinical or public health conclusions. By relaxing the core threshold, stable alignment sizes can be maintained up to and likely beyond 10,000 genomes, improving phylogenetic accuracy without compromising data integrity.

Furthermore, the strict-core approach poses challenges when generating whole-genome alignments from enrichment or capture-based methods, which are valuable for sequencing unculturable pathogens but often result in higher rates of missing data. These enrichment and capture-based methods are relatively new and are being used more frequently for bacteria, necessitating the establishment of best-practice data analysis techniques (6, 11, 60, 61). Similarly, missing data may arise when sequencing viral genomes, especially when genomes are sequenced directly from clinical samples. The inherent diversity of viruses and often degraded nature of clinical samples exacerbate the problem, making it more challenging to achieve complete sequencing coverage (62). As a result, adjustments to account for missing data have become more prevalent in viral and enrichment WGS data, driving the adoption of more adaptable analytical methods. For these datasets, a soft-core approach conserves data that are part of the clonal frame of evolution but sometimes fail to sequence (63). To address this issue, enrichment and viral studies have started to adopt soft-core methods using alternative tools, highlighting the need for more flexible approaches to preserve critical genomic information (6, 11).

For smaller datasets, particularly those with a strong temporal signal and limited geographic diversity, discarding SNP positions using a strict core threshold can remove valuable information. A soft-core approach can be valuable here and complement other tools such as *Gubbins* for recombination filtering. Our analysis showed that for recombination-unfiltered alignments, a 95% core threshold generally performs better (e.g., *S. pneumoniae* and *N. gonorrhoeae*) or similarly (e.g., *S. aureus* and *S.* Kentucky) to a 100% core threshold, making 95% the preferred threshold. For recombination-filtered alignments, the core threshold showed little impact on temporal signal values, except for a clear decline near 100%. Thresholds ranging from 50% to 95% performed similarly, suggesting that overly strict cores discard valuable data. For example, our recombination-filtered *N. gonorrhoeae* dataset lost over 90% of SNPs at a 100% core threshold. Overall, a 95% core threshold generally outperforms a 100% threshold.

We also note that phylogenetic trees constructed from a 95% core threshold alignment can result in better-resolved topologies than those from 100% core threshold alignments (64). This is because including sites with some gaps among samples provides more information to resolve topologies and estimate branch lengths among samples where a base is present. Subsequent analyses, including phylodynamic inference, can in turn benefit from better-resolved topologies. Including samples with some gaps may provide an avenue for a more inclusive analysis, as these samples may contain informative metadata that can contribute to a more accurate phylogenetic reconstruction. This includes patient travel history and sampling time information that improves phylogeographic and phylodynamic inference (65).

When comparing *Core-SNP-filter* to other tools with similar capabilities, the choice of tool has minimal impact for small datasets (e.g., <100 samples) as all tested tools are sufficiently fast and efficient. However, for large datasets (e.g., >1000 samples), especially those containing invariant sites, both time and RAM become critical factors, and *Core-SNP-filter* has the potential to scale to very large datasets (e.g., >100,000 samples). *Goalign* presents challenges due to its non-intuitive commands for excluding invariant sites and producing soft-core alignments, though it has functionality for processing alignments in ways that *Core-SNP-filter* cannot.

Our study underscores the advantages of using soft-core thresholds, particularly the 95% threshold, over strict 100% thresholds for bacterial genome datasets. The soft-core approach, achieved with *Core-SNP-filter,* preserves more informative sites, enhancing phylogenetic resolution and temporal signal, while being computationally efficient. Researchers and public health professionals should consider using a soft core in their analyses to retain available evolutionary signal and accurately interpret phylogenetic relationships and transmission patterns.

CONFLICTS OF INTEREST

No conflict of interest declared.

FUNDING INFORMATION

MLT is supported by an Australian Government Research Training Program Scholarship. DJI is supported by an NHMRC Investigator Grant (GNT1195210). TPS is supported by the National Health and Medical Research Council of Australia Investigator Grant (GNT1194325).

ETHICAL APPROVAL

No ethics was required for this study.

AUTHOR CONTRIBUTIONS

MLT, RRW, GT, and LAF designed the study. MLT and RRW performed all analyses. MLT and RRW prepared the figures. GT, DJI and TPS provided supervision and support. The first draft of the manuscript was written by MLT, being revised by all authors. All authors had full access to all study data. The corresponding author had final responsibility for the decision to submit for publication.

REFERENCES

1. Wooding S. Inferring Phylogenies. (0002-9297 (Print)).

2. Attwood SW, Hill SC, Aanensen DM, Connor TR, Pybus OG. Phylogenetic and phylodynamic approaches to understanding and combating the early SARS-CoV-2 pandemic. Nat Rev Genet. 2022;23(9):547-62.

3. Colijn C, Gardy J. Phylogenetic tree shapes resolve disease transmission patterns. Evolution, Medicine, and Public Health. 2014;2014(1):96-108.

4. Lane CR, Sherry NL, Porter AF, Duchene S, Horan K, Andersson P, et al. Genomics-informed responses in the elimination of COVID-19 in Victoria, Australia: an observational, genomic epidemiological study. The Lancet Public Health. 2021;6(8):e547-e56.

5. Seemann T, Lane CR, Sherry NL, Duchene S, Gonçalves da Silva A, Caly L, et al. Tracking the COVID-19 pandemic in Australia using genomics. Nat Commun. 2020;11(1):4376.

6. Taouk ML, Taiaroa G, Pasricha S, Herman S, Chow EPF, Azzatto F, et al. Characterisation of *Treponema pallidum* lineages within the contemporary syphilis outbreak in Australia: a genomic epidemiological analysis. The Lancet Microbe. 2022;3(6):e417-e26.

7. Williamson DA, Chow EPF, Gorrie CL, Seemann T, Ingle DJ, Higgins N, et al. Bridging of *Neisseria gonorrhoeae* lineages across sexual networks in the HIV pre-exposure prophylaxis era. Nature Communications. 2019;10(1):3988.

8. Croucher NJ, Page AJ, Connor TR, Delaney AJ, Keane JA, Bentley SD, et al. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. Nucleic Acids Research. 2014;43(3):e15-e.

9. Didelot X, Wilson DJ. ClonalFrameML: Efficient Inference of Recombination in Whole Bacterial Genomes. PLOS Computational Biology. 2015;11(2):e1004041.

10. Baines SL, Holt KE, Schultz MB, Seemann T, Howden BO, Jensen SO, et al. Convergent Adaptation in the Dominant Global Hospital Clone ST239 of Methicillin-Resistant *Staphylococcus aureus*. mBio. 2015;6(2):10.1128/mbio.00080-15.

11. Beale MA, Marks M, Cole MJ, Lee M-K, Pitt R, Ruis C, et al. Global phylogeny of *Treponema pallidum* lineages reveals recent expansion and spread of contemporary syphilis. Nature Microbiology. 2021;6(12):1549-60.

12. Golparian D, Harris SR, Sánchez-Busó L, Hoffmann S, Shafer WM, Bentley SD, et al. Genomic evolution of *Neisseria gonorrhoeae* since the preantibiotic era (1928–2013): antimicrobial use/misuse selects for resistance and drives evolution. BMC Genomics. 2020;21(1):116.

13. Gorrie CL, Mirčeta M, Wick RR, Judd LM, Lam MMC, Gomi R, et al. Genomic dissection of *Klebsiella pneumoniae* infections in hospital patients reveals insights into an opportunistic pathogen. Nature Communications. 2022;13(1):3017.

14. Hawkey J, Le Hello S, Doublet B, Granier SA, Hendriksen RS, Fricke WF, et al. Global phylogenomics of multidrug-resistant *Salmonella enterica* serotype Kentucky ST198. Microb Genom. 2019;5(7).

15. Ingle DJ, Ambrose RL, Baines SL, Duchene S, Gonçalves da Silva A, Lee DYJ, et al. Evolutionary dynamics of multidrug resistant *Salmonella enterica* serovar 4,[5],12:i:- in Australia. Nature Communications. 2021;12(1):4786.

16. Ingle DJ, Andersson P, Valcanis M, Barnden J, Silva AGd, Horan KA, et al. Prolonged Outbreak of Multidrug-Resistant *Shigella sonnei* Harboring bla sub CTX-M-27 in Victoria, Australia. Antimicrobial Agents and Chemotherapy. 2020;64(12):10.1128/aac.01518-20.

17. Thomas JC, Seby S, Abrams AJ, Cartee J, Lucking S, Vidyaprakash E, et al. Evidence of Recent Genomic Evolution in Gonococcal Strains With Decreased Susceptibility to Cephalosporins or Azithromycin in the United States, 2014-2016. J Infect Dis. 2019;220(2):294-305.

18. Satam H, Joshi K, Mangrolia U, Waghoo S, Zaidi G, Rawool S, et al. Next-Generation Sequencing Technology: Current Trends and Advancements. Biology (Basel). 2023;12(7).

19. Carey ME, Dyson ZA, Ingle DJ, Amir A, Aworh MK, Chattaway MA, et al. Global diversity and antimicrobial resistance of typhoid fever pathogens: Insights from a meta-analysis of 13,000 *Salmonella* Typhi genomes. eLife. 2023;12:e85867.

20. Harrison OB, Cehovin A, Skett J, Jolley KA, Massari P, Genco CA, et al. *Neisseria gonorrhoeae* Population Genomics: Use of the Gonococcal Core Genome to Improve Surveillance of Antimicrobial Resistance. J Infect Dis. 2020;222(11):1816-25.

21. Higgs C, Sherry NL, Seemann T, Horan K, Walpola H, Kinsella P, et al. Optimising genomic approaches for identifying vancomycin-resistant *Enterococcus faecium* transmission in healthcare settings. Nature Communications. 2022;13(1):509.

22. Mason LCE, Greig DR, Cowley LA, Partridge SR, Martinez E, Blackwell GA, et al. The evolution and international spread of extensively drug resistant *Shigella sonnei*. Nature Communications. 2023;14(1):1983.

23. Gorrie CL, Da Silva AG, Ingle DJ, Higgs C, Seemann T, Stinear TP, et al. Key parameters for genomics-based real-time detection and tracking of multidrug-resistant bacteria: a systematic analysis. The Lancet Microbe. 2021;2(11):e575-e83.

24. Fitzpatrick MA, Ozer EA, Hauser AR. Utility of Whole-Genome Sequencing in Characterizing Acinetobacter Epidemiology and Analyzing Hospital Outbreaks. J Clin Microbiol. 2016;54(3):593-612.

25. Giulieri SG, Guérillot R, Holmes NE, Baines SL, Hachani A, Hayes AS, et al. A statistical genomics framework to trace bacterial genomic predictors of clinical outcomes in *Staphylococcus aureus* bacteremia. Cell Rep. 2023;42(9):113069.

26. Argimón S, David S, Underwood A, Abrudan M, Wheeler NE, Kekre M, et al. Rapid Genomic Characterization and Global Surveillance of Klebsiella Using Pathogenwatch. Clin Infect Dis. 2021;73(Suppl_4):S325-s35.

27. Chung GT, Yoo JS, Oh HB, Lee YS, Cha SH, Kim SJ, et al. Complete genome sequence of *Neisseria gonorrhoeae* NCCP11945. J Bacteriol. 2008;190(17):6035-6.

28. Parkhill J, Dougan G, James KD, Thomson NR, Pickard D, Wain J, et al. Complete genome sequence of a multiple drug resistant *Salmonella enterica* serovar Typhi CT18. Nature. 2001;413(6858):848-52.

29. Alfsnes K, Eldholm V, Olsen AO, Brynildsrud OB, Bohlin J, Steinbakk M, et al. Genomic epidemiology and population structure of *Neisseria gonorrhoeae* in Norway, 2016-2017. Microb Genom. 2020;6(4).

30. Buckley C, Forde BM, Trembizki E, Lahra MM, Beatson SA, Whiley DM. Use of whole genome sequencing to investigate an increase in *Neisseria gonorrhoeae* infection among women in urban areas of Australia. Scientific Reports. 2018;8(1):1503.

31. Cehovin A, Harrison OB, Lewis SB, Ward PN, Ngetsa C, Graham SM, et al. Identification of Novel *Neisseria gonorrhoeae* Lineages Harboring Resistance Plasmids in Coastal Kenya. J Infect Dis. 2018;218(5):801-8.

32. De Silva D, Peters J, Cole K, Cole MJ, Cresswell F, Dean G, et al. Whole-genome sequencing to determine transmission of *Neisseria gonorrhoeae*: an observational study. Lancet Infect Dis. 2016;16(11):1295-303.

33. Demczuk W, Lynch T, Martin I, Domselaar GV, Graham M, Bharat A, et al. Whole-Genome Phylogenomic Heterogeneity of *Neisseria gonorrhoeae* Isolates with Decreased Cephalosporin Susceptibility Collected in Canada between 1989 and 2013. Journal of Clinical Microbiology. 2015;53(1):191-200.

34. Didelot X, Dordel J, Whittles LK, Collins C, Bilek N, Bishop CJ, et al. Genomic Analysis and Comparison of Two Gonorrhea Outbreaks. mBio. 2016;7(3).

35. Ezewudo MN, Joseph SJ, Castillo-Ramirez S, Dean D, Del Rio C, Didelot X, et al. Population structure of *Neisseria gonorrhoeae* based on whole genome data and its relationship with antibiotic resistance. PeerJ. 2015;3:e806.

36. Fifer H, Saunders J, Soni S, Sadiq ST, FitzGerald M. 2018 UK national guideline for the management of infection with *Neisseria gonorrhoeae*. Int J STD AIDS. 2020;31(1):4-15.

37. Grad YH, Harris SR, Kirkcaldy RD, Green AG, Marks DS, Bentley SD, et al. Genomic Epidemiology of Gonococcal Resistance to Extended-Spectrum Cephalosporins, Macrolides, and Fluoroquinolones in the United States, 2000-2013. J Infect Dis. 2016;214(10):1579-87.

38. Kwong JC, Chow EPF, Stevens K, Stinear TP, Seemann T, Fairley CK, et al. Whole-genome sequencing reveals transmission of gonococcal antibiotic resistance among men who have sex with men: an observational study. Sex Transm Infect. 2018;94(2):151-7.

39. Kwong JC, Gonçalves da Silva A, Dyet K, Williamson DA, Stinear TP, Howden BP, et al. NGMASTER:in silico multi-antigen sequence typing for *Neisseria gonorrhoeae*. Microb Genom. 2016;2(8):e000076.

40. Lan PT, Golparian D, Ringlander J, Van Hung L, Van Thuong N, Unemo M. Genomic analysis and antimicrobial resistance of *Neisseria gonorrhoeae* isolates from Vietnam in 2011 and 2015–16. Journal of Antimicrobial Chemotherapy. 2020;75(6):1432-8.

41. Lee RS, Seemann T, Heffernan H, Kwong JC, Gonçalves da Silva A, Carter GP, et al. Genomic epidemiology and antimicrobial resistance of *Neisseria gonorrhoeae* in New Zealand. Journal of Antimicrobial Chemotherapy. 2017;73(2):353-64.

42. Mortimer TD, Pathela P, Crawley A, Rakeman JL, Lin Y, Harris SR, et al. The Distribution and Spread of Susceptible and Resistant *Neisseria gonorrhoeae* Across Demographic Groups in a Major Metropolitan Center. Clin Infect Dis. 2021;73(9):e3146-e55.

43. Ryan L, Golparian D, Fennelly N, Rose L, Walsh P, Lawlor B, et al. Antimicrobial resistance and molecular epidemiology using whole-genome sequencing of *Neisseria gonorrhoeae* in Ireland, 2014-2016: focus on extended-spectrum cephalosporins and azithromycin. Eur J Clin Microbiol Infect Dis. 2018;37(9):1661-72.

44. Sánchez-Busó L, Golparian D, Corander J, Grad YH, Ohnishi M, Flemming R, et al. The impact of antimicrobials on gonococcal evolution. Nature Microbiology. 2019;4(11):1941-50.

45. Town K, Field N, Harris SR, Sánchez-Busó L, Cole MJ, Pitt R, et al. Phylogenomic analysis of *Neisseria gonorrhoeae* transmission to assess sexual mixing and HIV transmission risk in England: a cross-sectional, observational, whole-genome sequencing study. Lancet Infect Dis. 2020;20(4):478-86.

46. Wind CM, de Vries E, Schim van der Loeff MF, van Rooijen MS, van Dam AP, Demczuk WHB, et al. Decreased Azithromycin Susceptibility of *Neisseria gonorrhoeae* Isolates in Patients Recently Treated with Azithromycin. Clin Infect Dis. 2017;65(1):37-45.

47. Yahara K, Nakayama SI, Shimuta K, Lee KI, Morita M, Kawahata T, et al. Genomic surveillance of *Neisseria gonorrhoeae* to investigate the distribution and evolution of antimicrobial-resistance determinants and lineages. Microb Genom. 2018;4(8).

48. Croucher NJ, Harris SR, Fraser C, Quail MA, Burton J, van der Linden M, et al. Rapid Pneumococcal Evolution in Response to Clinical Interventions. Science. 2011;331(6016):430-4.

49. Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. Molecular Biology and Evolution. 2014;32(1):268-74.

50. Paradis E, Schliep K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics. 2018;35(3):526-8.

51. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics. 2009;25(15):1972-3.

52. Page AJ, Taylor B, Delaney AJ, Soares J, Seemann T, Keane JA, et al. SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. Microbial Genomics. 2016;2(4).

53. Lemoine F, Gascuel O. Gotree/Goalign: toolkit and Go API to facilitate the development of phylogenetic workflows. NAR Genomics and Bioinformatics. 2021;3(3).

54. Biek R, Pybus OG, Lloyd-Smith JO, Didelot X. Measurably evolving pathogens in the genomic era. Trends in Ecology & Evolution. 2015;30(6):306-13.

55. Featherstone LA, Zhang JM, Vaughan TG, Duchene S. Epidemiological inference from pathogen genomes: A review of phylodynamic models and applications. Virus Evolution. 2022;8(1).

56. Hennart M, Guglielmini J, Bridel S, Maiden MCJ, Jolley KA, Criscuolo A, et al. A Dual Barcoding Approach to Bacterial Strain Nomenclature: Genomic Taxonomy of *Klebsiella pneumoniae* Strains. Molecular Biology and Evolution. 2022;39(7).

57. Jajou R, Kohl TA, Walker T, Norman A, Cirillo DM, Tagliani E, et al. Towards standardisation: comparison of five whole genome sequencing (WGS) analysis pipelines for detection of epidemiologically linked tuberculosis cases. Eurosurveillance. 2019;24(50):1900130.

58. Mixão V, Pinto M, Sobral D, Di Pasquale A, Gomes JP, Borges V. ReporTree: a surveillance-oriented tool to strengthen the linkage between pathogen genetic clusters and epidemiological data. Genome Medicine. 2023;15(1):43.

59. Stimson J, Gardy J, Mathema B, Crudu V, Cohen T, Colijn C. Beyond the SNP Threshold: Identifying Outbreak Clusters Using Inferred Transmissions. Molecular Biology and Evolution. 2019;36(3):587-603.

60. Depledge DP, Palser AL, Watson SJ, Lai IY-C, Gray ER, Grant P, et al. Specific Capture and Whole-Genome Sequencing of Viruses from Clinical Samples. PLOS ONE. 2011;6(11):e27805.

61. Mee PT, Buultjens AH, Oliver J, Brown K, Crowder JC, Porter JL, et al. Mosquitoes provide a transmission route between possums and humans for Buruli ulcer in southeastern Australia. Nature Microbiology. 2024;9(2):377-89.

62. Martin GE, Taiaroa G, Taouk ML, Savic I, O'Keefe J, Quach R, et al. Maintaining genomic surveillance using whole-genome sequencing of SARS-CoV-2 from rapid antigen test devices. The Lancet Infectious Diseases. 2022;22(10):1417-8.

63. Taouk ML, Steinig E, Taiaroa G, Savic I, Tran T, Higgins N, et al. Intra- and interhost genomic diversity of monkeypox virus. Journal of Medical Virology. 2023;95(8):e29029.

64. Felsenstein J. Evolutionary trees from DNA sequences: A maximum likelihood approach. Journal of Molecular Evolution. 1981;17(6):368-76.

65. Porter AF, Featherstone L, Lane CR, Sherry NL, Nolan ML, Lister D, et al. The importance of utilizing travel history metadata for informative phylogeographical inferences: a case study of early SARS-CoV-2 introductions into Australia. Microbial Genomics. 2023;9(8).