

Integrated population clustering and genomic epidemiology with PopPIPE

Martin P. McHugh^{1,2}, Samuel T. Horsfield³, Johanna von Wachsman³, Jacqueline Toussaint³, Kerry A. Pettigrew^{1,2,†}, Elzbieta Czarniak¹, Thomas J. Evans⁴, Alistair Leanord^{4,5}, Luke Tysall¹, Stephen H. Gillespie², Kate E. Templeton¹, Matthew T. G. Holden², Nicholas J. Croucher⁶ and John A. Lees^{3,*}

Abstract

Genetic distances between bacterial DNA sequences can be used to cluster populations into closely related subpopulations and as an additional source of information when detecting possible transmission events. Due to their variable gene content and order, reference-free methods offer more sensitive detection of genetic differences, especially among closely related samples found in outbreaks. However, across longer genetic distances, frequent recombination can make calculation and interpretation of these differences more challenging, requiring significant bioinformatic expertise and manual intervention during the analysis process. Here, we present a **Population analysis PIPEline** (PopPIPE) which combines rapid reference-free genome analysis methods to analyse bacterial genomes across these two scales, splitting whole populations into subclusters and detecting plausible transmission events within closely related clusters. We use k-mer sketching to split populations into strains, followed by split k-mer analysis and recombination removal to create alignments and subclusters within these strains. We first show that this approach creates high-quality subclusters on a population-wide dataset of *Streptococcus pneumoniae*. When applied to nosocomial vancomycin-resistant *Enterococcus faecium* samples, PopPIPE finds transmission clusters that are more epidemiologically plausible than core genome or multilocus sequence typing (MLST) approaches. Our pipeline is rapid and reproducible, creates interactive visualizations and can easily be reconfigured and re-run on new datasets. Therefore, PopPIPE provides a user-friendly pipeline for analyses spanning species-wide clustering to outbreak investigations.

Impact Statement

As time passes, bacterial genomes accumulate small changes in their sequence due to mutations, or larger changes in their content due to horizontal gene transfer. Using their genome sequences, it is possible to use phylogenetics to work out the most likely order in which these changes happened, and how long they took to happen. Then, one can estimate the time that separates any two bacterial samples – if it is short, then they may have been directly transmitted or acquired from the same source; but if it is long, they must have been acquired separately. This information can be used to determine transmission chains, in conjunction with dates and locations of infections. Understanding transmission chains enables targeted infection control measures. However, correctly calculating the genetic evidence for transmission is made difficult by correctly distin-

[Continued on next page]

Received 13 December 2024; Accepted 22 March 2025; Published 28 April 2025

Author affiliations: ¹Medical Microbiology, Department of Laboratory Medicine, Royal Infirmary of Edinburgh, Edinburgh EH16 4SA, UK; ²Division of Infection and Global Health, University of St Andrews, St Andrews KY16 9AJ, UK; ³European Molecular Biology Laboratory, European Bioinformatics Institute EMBL-EBI, Hinxton CB10 1SD, UK; ⁴School of Infection and Immunity, University of Glasgow, Glasgow G12 8QQ, UK; ⁵Scottish Microbiology Reference Laboratories, Glasgow Royal Infirmary, Glasgow G4 0SF, UK; ⁶MRC Centre for Global Infectious Disease Analysis, School of Public Health, Imperial College London, London SW7 2AZ, UK.

***Correspondence:** John A. Lees, jlees@ebi.ac.uk

Keywords: clustering; genomic epidemiology; pipelines; transmission.

Abbreviations: cgMLST, core genome MLST; MLST, Multilocus sequence typing; PopPIPE, population analysis PIPEline; SKA, Split k-mer analysis; VREfm, Vancomycin-resistant *Enterococcus faecium*; WGS, whole-genome sequencing.

†Present address: School of Health and Life Sciences, Teesside University, Middlesbrough TS1 3BX, UK.

001404 © 2025 The Authors



This is an open-access article distributed under the terms of the Creative Commons Attribution License. This article was made open access via a Publish and Read agreement between the Microbiology Society and the corresponding author's institution.

guishing different types of sequence changes, dealing with large amounts of genome data and the need to use multiple complex bioinformatic tools. We addressed this gap by creating a computational workflow, PopPIPE, which automates the process of detecting possible transmissions using genome sequences. PopPIPE applies state-of-the-art tools and is fast and easy to run – making this technology available to a wider audience of researchers.

DATA SUMMARY

The code for this pipeline is available at <https://github.com/bacpop/PopPIPE> and as a docker image <https://hub.docker.com/r/poppunk/poppipe>. Raw sequencing reads for *Enterococcus faecium* isolates have been deposited at the NCBI under BioProject accession no. PRJNA997588. Results of the subcluster analysis have been deposited on figshare <https://dx.doi.org/10.6084/m9.figshare.28429574>. Input files and results of the transmission analysis have been deposited on figshare <https://dx.doi.org/10.6084/m9.figshare.28495571>.

INTRODUCTION

Routine genomic surveillance and targeted sequencing of bacterial pathogens have become central tenets of infectious disease epidemiology. Analysis of genomes from bacterial pathogens, at its simplest, will give information on the species and sequence cluster (also called strain, sequence type, lineage or clade) of the sequenced isolate. Typically, these are two compatible hierarchical definitions, often made at 95% and 99.5% sequence identity, respectively [1–3]. Genetically determined species largely agree with traditional prokaryotic taxonomy and determine microbiologically and clinically important differences between infectious agents [4]. Strain designations sometimes have useful correlation with phenotypes (e.g. virulence and resistance) due to inherited genetic elements within the clades they represent.

Strain designations are powerful at *ruling out* recent transmission between samples, as the genetic distance between two strains frequently represents hundreds or thousands of years of evolution [5]. A further level of sub-clustering within strains, which we refer to as lineages, can be useful for making this ruling out more sensitive (without loss of precision). However, genetic distances alone are less powerful at *ruling in* transmission [6] and, for this purpose, are best combined with epidemiological data on sampling time [7] and place [8]. To bring these insights from complex models into epidemiological practice also requires judicious use of suitable interactive visualizations [9, 10].

Integrated genomic and epidemiological analysis and visualization of multi-strain bacterial populations are typically complicated due to the fact that across the entire species a vertical evolutionary history cannot be accurately reconstructed from contemporary genomes alone [11]. If this is not accounted for, genetic branch lengths are likely to be overestimated, causing significant errors in sub-clustering and transmission analyses [12].

Furthermore, selecting correct and efficient methods from a wide array of possibilities, connecting them together and making them reproducible is a challenge. For epidemiological analysis using genomes, particularly if this analysis is going to be used to inform interventions or policy changes, it is critical that results are reliable and reproducible. In some cases where external validation of pipelines is required for use in public health, their constituent programmes and parameters must remain stable so that their evaluation performance on validation datasets is unchanged [13]. Even for experienced bioinformaticians, best practices from multiple pieces of software can be hard to follow or adapt exactly to the use case in hand. These issues can be helped by pipeline managers such as Snakemake and Nextflow [14, 15]. Using these managers, pipelines can run with set programmes and versions and transparently set configurations that can be applied across different datasets.

We, therefore, designed the **Population analysis PIPEline** (PopPIPE) to automate the complex orchestration of running many bioinformatic tools for sub-clustering and transmission analysis across bacterial populations with many strains. By using sensible, but easily adjustable, defaults and efficient choices of bioinformatic software tools, the pipeline runs rapidly, requires minimal user input and produces reliable and reproducible genomic inferences. Our pipeline has three modes: sub-clustering, visualization and transmission detection. We demonstrate its operation on a population-wide dataset of *Streptococcus pneumoniae* and a hospital outbreak of vancomycin-resistant *Enterococcus faecium* (VREfm).

THEORY AND IMPLEMENTATION

Phylogenetics and genomic-based transmission detection rely on an accurate clonal frame where vertical evolution can be inferred as a series of sequential events along a tree. If recombination events are not removed prior to these steps, branch lengths will be overestimated [12] and transmission events/clusters will be missed. Methods exist to remove these recombination events (e.g. gubbins [16] and ClonalFrameML [17]) but are only applicable within populations of limited diversity sharing a recent common ancestor.

Many bacterial species have populations that maintain a genetic structure of different strains (also called sequence types). Analogous to an average nucleotide identity of 95% separating most species, an average nucleotide identity of around 99% separates most strains [18–20]. However, it is often challenging to reconstruct any valid vertical evolutionary history across these strains due to frequent and repeated overwriting of sequences resulting from homologous recombination and the movement of mobile genetic elements [11]. Although gene-based methods such as FastGEAR [21] can detect ancestral recombination events, these cannot capture the whole-genome resolution needed to accurately distinguish closely related samples in an outbreak setting.

These are complex evolutionary considerations that may not always be easy to address or within the scope of a genomic epidemiology solution. We, therefore, present a pragmatic and general solution to this issue: a population is first subdivided into genetic clusters *between which* transmission is completely implausible, and *within which* a recent common ancestor exists. Then, in every such cluster, recombination can be detected and removed, leaving a clonal frame from which phylogenetics and transmission detection can be performed. Specifically, within each cluster, we aim to produce the following useful components for genomic epidemiology analyses:

- Genetic subclusters, defined hierarchically with multiple nested levels of further subclusters.
- A maximum-likelihood phylogeny from whole-genome alignment.
- A recombination removed, dated phylogeny.
- Inferred transmission events based on this phylogeny.
- Interactive visualization of the genetic relationships and subclusters for the whole dataset.

We refer to these subclusters as lineages, due to analogous usage in the non-multi-strain *Mycobacterium tuberculosis* [22] and viral pathogens [23].

We implemented this process in PopPIPE (v1.1.0), a Snakemake (v7.8.5) pipeline [14]. The pipeline can, therefore, continue failed or paused analysis from where it left off, rerun the same analysis on new data completely consistently with previous runs and automatically take advantage of multicore systems or common job submission systems. Our pipeline is designed to be run following a PopPUNK strain-typing analysis [2], which is first used to split a population into strains. The data needed as input is:

- A user's self-made PopPUNK database, or an existing PopPUNK database if stable strain nomenclature is desired (<https://www.bacpop.org/poppunk/>).
- Assemblies (if available, preferred for speed) or Illumina sequence reads for all samples included.
- If transmission is to be inferred, a file listing the dates all samples were taken.
- A config file, listing file paths of these inputs and defining parameters for tools in the pipeline.

As PopPIPE depends on an initial PopPUNK clustering, all downstream results will be dependent on the quality of this primary clustering analysis. We, therefore, recommend checking the quality of these strain designations before running PopPIPE or using a pre-built database model fit. PopPUNK has been shown to work for the most commonly sequenced bacterial species. For lower diversity species, for example *M. tuberculosis*, users may opt to choose a PopPUNK model that creates only one large cluster (by setting a large `--threshold` or otherwise), and allowing PopPIPE to do all further subclustering.

The steps of the default pipeline which makes subclusters of the data are as follows (Fig. 1):

- (1) Partition assembly/read files into their PopPUNK (v2.7.0) [2] strains. Steps 2–6 are then carried out in parallel across each strain.
- (2) Use `pp-sketchlib` (v2.1.3) to calculate core and accessory distances within each strain, using the same parameters as the PopPUNK database.
- (3) Use the core distances with `RapidNJ` (v2.3.2) [24] to make a neighbour-joining tree.
- (4) Use `SKA2` (v0.3.9) [20] to generate reference-free alignments using `ska align`.
- (5) Use `IQ-TREE` (v2.0.3) [25] to generate a maximum-likelihood phylogeny using this alignment and the neighbour-joining tree as a starting point.
- (6) Use `Fastbaps` (v1.0.5) [26] to generate subclusters/lineages, in the mode where subclusters are necessarily partitions of the phylogeny (increasing speed and eliminating polyphyletic clusters).

To make an interactive visualization of the subclusters in `microreact` [27], two further steps are run:

- (7) Use `Mandrake` (v1.2.1) [28] to create an embedding visualization of the accessory distances.
- (8) Create an overall tree by grafting the maximum-likelihood trees for subclusters to their matching nodes, rescaling branch lengths to match the neighbour-joining tree and midpoint rooting maximum-likelihood trees. We note that this tree will not maximize the phylogenetic likelihood – it is only intended as a convenient visualization of the entire dataset, which is compatible with existing tools.

This creates a `.microreact` file as output, and if an API key is provided, it will automatically create a permanent URL hosting the visualization.

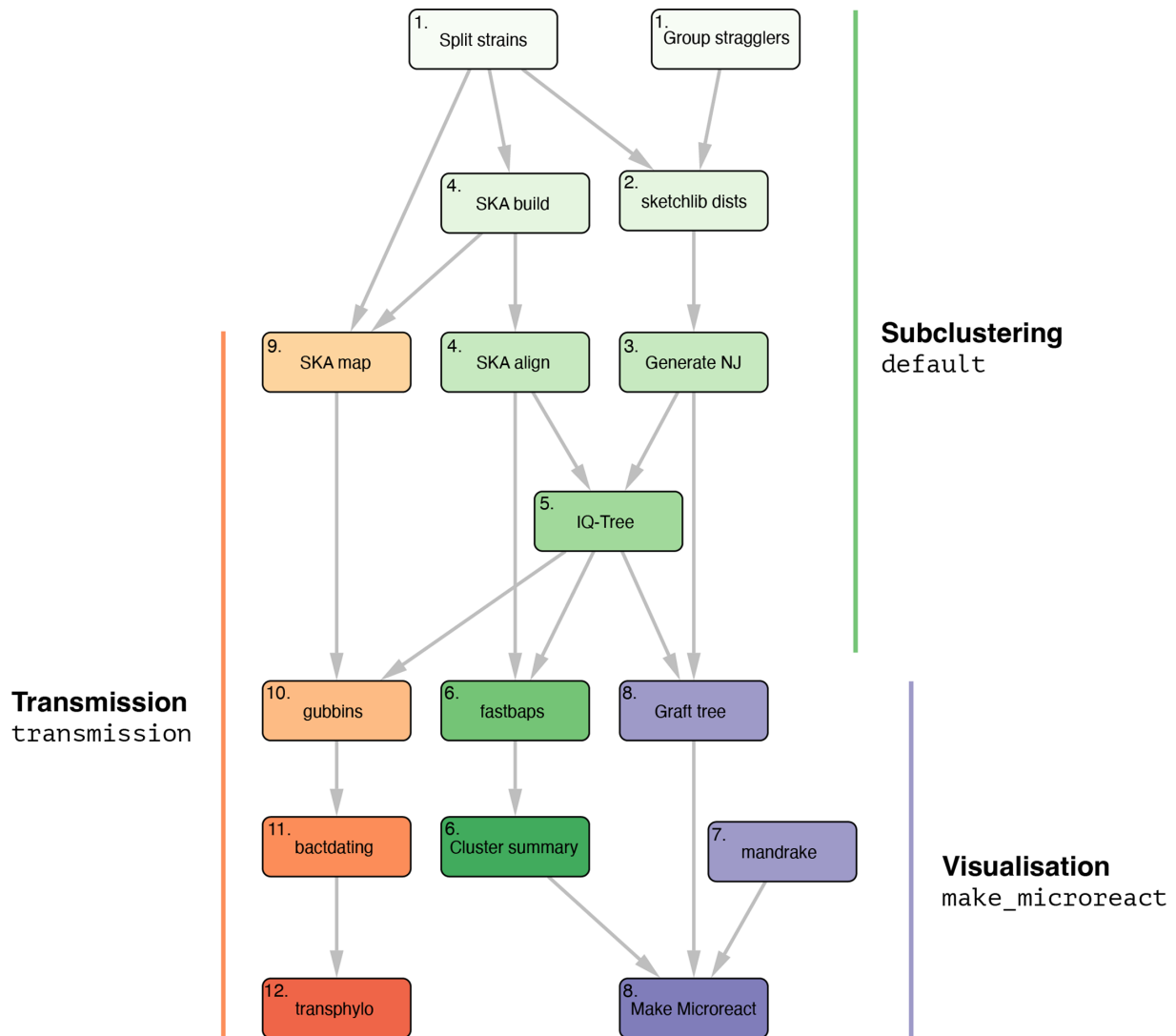


Fig. 1. Graph representing steps in PopPIPE. Nodes are snakemake rules, usually running a piece of software, edges are dependencies where a step will only start running when all its dependencies have been completed. Nodes are annotated with the step number where they are fully described in the text. Green nodes are run for the *default* target of generating sub-clusters. Purple nodes are run when visualization is being made with the *make_microreact* target. Orange nodes are run when transmission detection is run with the *transmission* target.

To infer transmission trees, four further steps are run:

- (9) SKA2 generates alignments with *ska map* and a reference genome from the strain so SNPs form correct windows on the physical chromosome.
- (10) Use Gubbins (v3.1.0) [16] to remove recombination.
- (11) Use BactDating (v1.1) [29] with sampling times to infer timed trees from recombination-purged phylogenies.
- (12) Use Transphylo (v1.4.8) [7] to infer possible transmission events on these timed trees.

SUB-CLUSTERING AND VISUALIZATION OF *S. PNEUMONIAE*

We demonstrated PopPIPE on 616 previously published *S. pneumoniae* genomes [19, 30]. PopPUNK database construction and clustering took 2 min using four threads. Using PopPIPE to create a visualization with a phylogeny and subclusters took an additional 28 min using four threads. The majority of this time was spent on maximum-likelihood tree inference – 50% of the total was spent on inferring the tree of largest strain (98 samples). Of the 62 inferred strains, 28 had a sufficient number of members to subcluster (≥ 6), producing 101 first-level subclusters and 157 second-level subclusters (Fig. 2). Running the transmission pipeline took an additional 52 min, the majority of the time being spent on recombination removal.

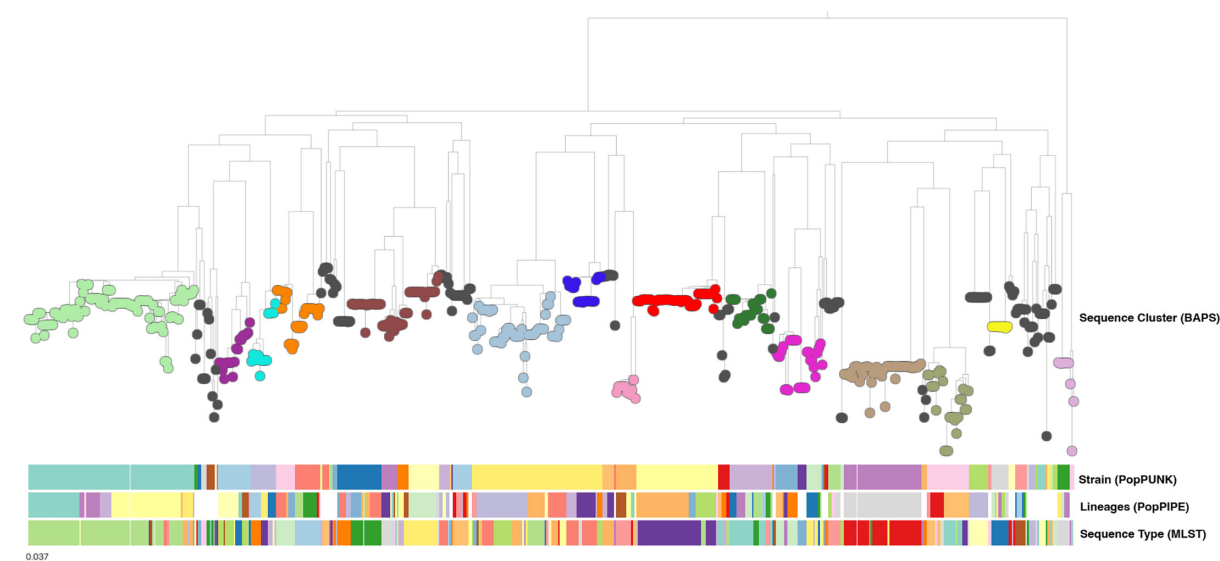


Fig. 2. Cluster results from the original core genome analysis with BAPS (colour of nodes), top-level strains from PopPUNK (first metadata block row), first-level subclusters (lineages) from PopPIPE (second metadata block row) and sequence types from MLST (third metadata block row). Some lineages are missing as the clusters are too small to cluster further. The black sequence cluster is the 'bin' cluster of grouped rare clusters.

The results of this pipeline run can also be found online (<https://microreact.org/project/sparc-poppipe>) and in Fig. S1, available in the online version of this article. We compared this to the original analysis of this data (also available on microreact <https://microreact.org/project/NJwviE7F>), which used a concatenated core gene alignment followed by maximum-likelihood tree construction [19]. Qualitatively similar results were found for tree topology and clusters (Fig. 2). A total of 62 strains were seen with the PopPIPE analysis, compared with 16 sequence clusters in the original analysis, which have a mutual information score of 2.49 (compared with 0.53 when randomly shuffled). Ignoring the 'bin' cluster of grouped rare clusters, there are 23 strains from PopPIPE and 15 sequence clusters, with the same mutual information score. The mutual information score of the first level subclusters was 3.38 with the sequence clusters and 3.74 with sequence types (compared with 1.48 when randomly shuffled), showing that the subclusters more closely correspond to MLST classifications.

A comparison of the tree topology in phylo.io [31] showed differences in the ancestral branches of the tree topology, as expected due to unresolvable recombination events, but generally close concordance near the tips within individual strains (Figs S2 and S3). The rooting and topology used in PopPIPE is not optimized, so it gives a worse tree. However, by construction, the subclusters/lineages necessarily matched the tree and its resolution, whereas sequence types offered inconsistent resolution across the tree as observed in [32]. As a more quantitative comparison, we also compared the topology of all of the subtrees for the 28 strains with at least six samples in the PopPIPE data (and hence, the full transmission pipeline could be run on them) using the mutual clustering information metric [33]. Eight were identical and 14 had only minor topology changes (mutual clustering in <0.75 and confirmed visually as in [34]). The most divergent was strain 4 (GPSC59; ST558; serotype 35B [35]) with a mutual clustering information of 0.68. This subtree has long terminal branches, a signal typical of unremoved recombination. Using the final subtree from PopPIPE, after recombination has been removed, is distant from both the PopPIPE tree and the original tree (which also did not have recombination removed), potentially more accurately representing vertical evolutionary history in this strain.

TRANSMISSION DETECTION IN VANCOMYCIN-RESISTANT *E. FAECIUM*

Several recent studies have also shown the utility of alignment-free clustering for investigation of *E. faecium* transmission [36–39]. Higgs *et al.* [36] report that SKA was more suitable than reference-based mapping to infer patient-to-patient transmission of *E. faecium* and recommend the use of SKA for this task going forward. Maechler *et al.* [37] compared clustering with core genome MLST (cgMLST), core genome SNPs and SKA with similar findings to those reported here – SKA generated more clusters with fewer genomes per cluster, but higher agreement with epidemiological data than both cgMLST and core SNPs. Both Maechler *et al.* [37] and Rath *et al.* [39] showed that SKA is suitable for *E. faecium* clustering in nosocomial settings and helps to rule in and rule out possible transmission links in combination with epidemiological data.

We, therefore, demonstrated the use of SKA within PopPIPE to subcluster a local outbreak dataset. VREfm readily spreads in healthcare settings leading to challenging outbreaks. Whole-genome sequencing (WGS) is increasingly used to identify putative transmission and target infection control measures. Short-read Illumina WGS was performed on 87 VREfm isolates from 84

Table 1. Patient epidemiological linkage within VREfm genomic clusters

Epidemiological linkage	PopPIPE [<i>n</i> =159] % (95% CI)	Core genome SNP [<i>n</i> =228] % (95% CI)
Same ward, same time	30.2 (23.1–37.3)	25.9 (20.2–31.6)
Same ward, stay 1–28 days apart	12.6 (7.4–17.8)	14.1 (9.6–18.6)
Different ward, same time	26.5 (19.6–33.4)	22.0 (16.6–27.4)
Different ward, stay 1–28 days apart	27.7 (20.7–34.7)	32.9 (26.8–39.0)
No link	3.2 (0.5–5.9)	5.3 (2.4–8.2)

CI, confidence interval; n, comparison pairs; SNP, single nucleotide polymorphism; VREfm, vancomycin resistant *Enterococcus faecium*.

patients included in an outbreak investigation across six orthopaedic inpatient wards. Adapters and low-quality regions were removed with Trimmomatic (v0.39) [40], and then the trimmed reads were assembled with SPAdes (v3.15.5) [41] using the *--isolate* flag. PopPUNK was used to create a database and subsequently cluster the assemblies. PopPIPE was then used to subcluster and infer transmission trees with transphylo. PopPIPE clustering was compared with reference-based core genome SNP clustering. Briefly, trimmed reads were mapped to the Aus0004 reference genome (accession no. CP003351) and mobile genetic elements masked using Snippy (v4.6.0) (<https://github.com/tseemann/snippy>), recombination was identified with Gubbins [16] and masked from the alignment, SNPs were extracted using SNP-sites (v2.4.0) [42], and pairwise SNP counts were calculated. Finally, core genome SNP clusters were identified with a threshold of ≤ 3 SNPs, as this was previously identified as the maximum diversity seen within individual patients [43]. Patient stay metadata was retrieved from electronic health records.

PopPIPE assigned 77 (88.6%) genomes into 20 clusters, while core genome SNPs grouped 74 (85.1%) genomes into 14 clusters. PopPIPE and core genome SNP grouping agreed in 68 (78.2%) cases. Within PopPIPE clusters, 30% of patient pairs were on the same ward at the same time and 3% had no epidemiological link, compared with 26% and 5% in core genome SNP clusters (Table 1). VREfm PopPIPE cluster 1-4-5 contained nine patients and was further examined as an example of transmission inference (Figs 3 and S4). The transmission tree suggested P1 transmitted to P4, both patients shared time and place in the hospital. P1 was also inferred to have transmitted to P8, who was on the same ward but at a different time; P34, who was on a different ward at the same time; and P30, who was epidemiologically unlinked. P8 was inferred to have transmitted to P19 and P21, all of whom were patients on the same ward at the same time. P4 was inferred to have transmitted to P16; these patients had no time or ward overlap in the hospital. P57 tested positive on screening at a pre-admission clinic, with no contact with the hospital in the preceding 6 months. This case was not linked to direct transmission in the transmission tree, which suggests pre-existing carriage from an unidentified source. TransPhylo was useful in ruling out possible direct transmission (P57) and for indicating likely linked patients (P1/P4 and P8/P19/P21), but for other cases, the inference should be examined with epidemiological data before considering direct transmission links. This reflects the challenges in inferring direct transmission, including the uncertain serial interval for *E. faecium*, long asymptomatic carriage, potential hospital environmental reservoirs and unsampled hosts.

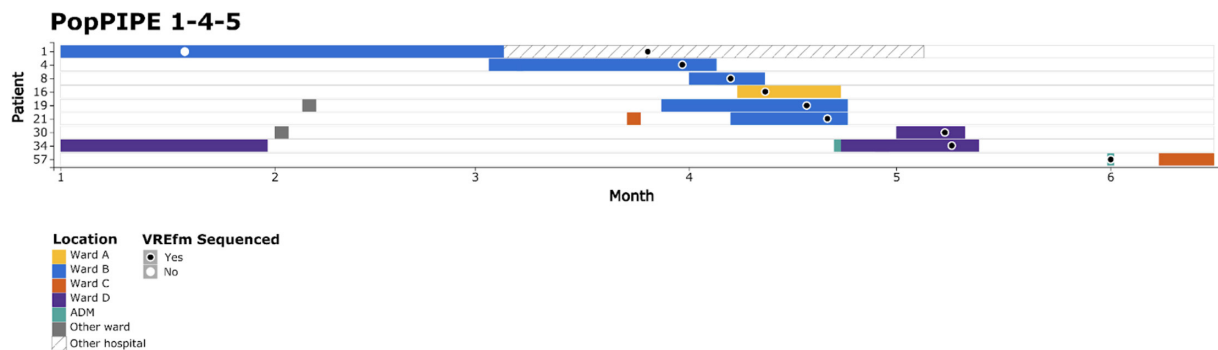


Fig. 3. Patient timelines for PopPIPE cluster 1-4-5. Each patient is represented by a row, with month on the x-axis. Hospital visits are plotted as coloured boxes, coloured by the ward, isolates are represented by circles and coloured based on whether the isolate was sequenced or not. ADM, pre-admission clinic. Figure generated with HALviz (v0.3) [50].

DISCUSSION

Recent computational methods use reference-free approaches to compare bacterial genomes more quickly and sensitively than traditional reference-based approaches. Sketching methods such as PopPUNK use variable length k -mers to rapidly calculate genetic distances across entire species and produce epidemiological clusters that divide populations into strains (usually separated by at least decades of evolution). SKA is a variant calling method which uses split k -mers to rapidly find specific variants within these strains. Combining sketching methods to split a population into first-level clusters, with fast variant calling within these clusters to obtain further subclusters, is an appealing approach to work with large sets of bacterial genomes. This combination of methods breaks analysis into manageable pieces which correspond to subpopulations where the vertical inheritance can be reconstructed with a degree of accuracy. Such methods are particularly needed in an era where millions of publicly available genomes are available to give context to population analyses [44], and where thousands of new genomes are sequenced every day [45].

By building on these underlying reference-free methods, PopPIPE represents a fast and reproducible method for the fundamental steps of population genetics using sets of genomes. What was previously the analysis underpinning an entire research project [19] can be completed automatically in under an hour, with the option to rerun or restart the same analysis with different parameters or be reproduced exactly by other users.

Genetic distances are also a potentially powerful addition to infectious disease epidemiology methods, supplementing traditional 'line lists' with case times and locations with an additional data source. Though genetic distances alone cannot typically resolve who-infected-whom [46], the combination of plausible evolutionary time, sampling time and physical proximity is a sensitive and specific way to detect transmission chains [8]. To obtain the best resolution on genetic distances in bacterial pathogens requires a reference-free and pangenome-wide approach [47], as traditional reference-based read-mapping methods will suffer from biases leading to decreased sensitivity and specificity [48]. Although SKA, the underlying reference-free method, is a fast and sensitive method to determine between sample genetic distances, it does not infer likely transmissions directly. Additionally, to achieve precision across a population requires careful prior clustering and recombination removal. We showed here that automating these steps in PopPIPE led to more precise and epidemiologically plausible transmission clusters than a core genome analysis.

We only tested PopPIPE using short-read Illumina data, but long-read data such as Oxford Nanopore or PacBio is increasingly attractive in outbreak investigation due to its portability and potential to resolve more variants in closely related outbreaks. PopPIPE could be used directly with long-read assemblies, but using long-reads directly is not supported as they cannot currently be error-corrected by various steps of the pipeline – therefore, users should first assemble their data before using PopPIPE. We would expect that more complete assemblies would give a more accurate distance estimation of closely related isolates, potentially improving branch length and transmission estimates. PopPIPE transmission is SNP-based, so it could not automatically take advantage of indels and structural variants that may be available in this data, but future versions could also include the *skalo* [49] method to attempt to include some of this information. Appropriate partition models for the phylogeny would also be needed, although these are not guaranteed to improve resolution [34].

We designed PopPIPE as an accessible route for researchers from all bioinformatic backgrounds to achieve high-quality SKA-based sub-clustering and transmission analysis. PopPIPE brings together a multitude of modern tools, initially designed to be run individually, together, enabling reference-free epidemiology. PopPIPE is, therefore, well placed for routine outbreak investigations given its speed, built-in visualization options and ease of repeating reproducible analyses with new data.

Funding information

S.T.H., J.v.W., J.T. and J.A.L. were supported by the European Molecular Biology Laboratory. S.T.H. and J.A.L. were additionally supported by BBSRC (reference BB/Y513805/1). N.J.C. was supported by the UK Medical Research Council and Department for International Development (grants MR/R015600/1 and MR/T016434/1). N.J.C. acknowledges funding from the MRC Centre for Global Infectious Disease Analysis (reference MR/X020258/1 and MR/T016434/1), funded by the UK Medical Research Council (MRC). This UK funded award is carried out in the frame of the Global Health EDCTP3 Joint Undertaking. M.P.M., K.A.P., T.J.E., A.L., S.H.G. and M.T.G.H. were supported by the Chief Scientist Office (Scotland) (grant number SIRN/10). Analysis of *E. faecium* data was performed on the Crop Diversity Bioinformatics HPC hosted by the James Hutton Institute and funded by BBSRC (reference BB/S019669/1).

Acknowledgements

We would like to thank Dr Donald Inverarity for assistance with *Enterococcus faecium* outbreak data.

Author contributions

Conceptualization: M.P.M., N.J.C., J.A.L.; Data curation: M.P.M., K.A.P., E.C., D.I., L.T., N.J.C.; Formal analysis: M.P.M., J.A.L.; Funding acquisition: T.J.E., A.L., S.H.G., M.T.G.H., N.J.C., J.A.L.; Methodology: J.A.L., J.v.W.; Software: S.T.H., J.v.W., J.T., J.A.L.; Resources: M.P.M., J.A.L.; Supervision: S.H.G., K.E.T., M.T.G.H., N.J.C., J.A.L.; Validation: J.T., N.J.C.; Visualization: M.P.M., J.A.L.; Writing – original draft: M.P.M., J.A.L.; Writing – review and editing: all authors.

Conflicts of interest

The authors declare that there are no conflicts of interest.

Ethical statement

Access to *E. faecium* isolates as excess diagnostic material was approved by the NHS Scotland BioRepository Network (reference no. TR000126). Access to enhanced patient metadata was approved by the NHS Lothian Caldicott Guardian (reference no. 1690).

References

- Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat Commun* 2018;9:5114.
- Lees JA, Harris SR, Tonkin-Hill G, Gladstone RA, Lo SW, et al. Fast and flexible bacterial genomic epidemiology with PopPUNK. *Genome Res* 2019;29:304–316.
- Rodriguez-R LM, Conrad RE, Viver T, Feistel DJ, Lindner BG, et al. An ANI gap within bacterial species that advances the definitions of intra-species units. *mBio* 2024;15:e0269623.
- Hedlund BP, Chuvochina M, Hugenholtz P, Konstantinidis KT, Murray AE, et al. SeqCode: a nomenclatural code for prokaryotes described from sequence data. *Nat Microbiol* 2022;7:1702–1708.
- Duchêne S, Holt KE, Weill F-X, Le Hello S, Hawkey J, et al. Genome-scale rates of evolutionary change in bacteria. *Microb Genom* 2016;2:e000094.
- Wymant C, Hall M, Ratmann O, Bonsall D, Golubchik T, et al. PHYLOSCANNER: Inferring transmission from within- and between-host pathogen genetic diversity. *Mol Biol Evol* 2018;35:719–733.
- Didelot X, Fraser C, Gardy J, Colijn C. Genomic infectious disease epidemiology in partially sampled and ongoing outbreaks. *Mol Biol Evol* 2017;34:997–1007.
- Cori A, Nouvellet P, Garske T, Bourhy H, Nakouné E, et al. A graph-based evidence synthesis approach to detecting outbreak clusters: an application to dog rabies. *PLoS Comput Biol* 2018;14:e1006554.
- Smith CM, Allen DJ, Nawaz S, Kozlakidis Z, Nastouli E, et al. An interactive data visualisation application to investigate nosocomial transmission of infections. *Wellcome Open Res* 2019;4:100.
- Nagraj VP, Randhawa N, Campbell F, Crellen T, Sudre B, et al. epicontacts: handling, visualisation and analysis of epidemiological contacts. *F1000Res* 2018;7:566.
- Sakoparnig T, Field C, van Nimwegen E. Whole genome phylogenies reflect the distributions of recombination rates for many bacterial species. *Elife* 2021;10:e65366. Epub ahead of print 8 January 2021.
- Hedge J, Wilson DJ. Bacterial phylogenetic reconstruction from whole genomes is robust to recombination but demographic inference is not. *mBio* 2014;5:e02158.
- Sherry NL, Horan KA, Ballard SA, Gonçalves da Silva A, Gorrie CL, et al. An ISO-certified genomics workflow for identification and surveillance of antimicrobial resistance. *Nat Commun* 2023;14:60.
- Mölder F, Jablonski KP, Letcher B, Hall MB, Tomkins-Tinch CH, et al. Sustainable data analysis with Snakemake. *F1000Res* 2021;10:33.
- Di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, et al. Nextflow enables reproducible computational workflows. *Nat Biotechnol* 2017;35:316–319.
- Croucher NJ, Page AJ, Connor TR, Delaney AJ, Keane JA, et al. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Res* 2015;43:e15.
- Didelot X, Wilson DJ. ClonalFrameML: efficient inference of recombination in whole bacterial genomes. *PLoS Comput Biol* 2015;11:e1004041.
- Hennart M, Guglielmini J, Bridel S, Maiden MCJ, Jolley KA, et al. A dual barcoding approach to bacterial strain nomenclature: genomic taxonomy of *Klebsiella pneumoniae* strains. *Mol Biol Evol* 2022;39:msac135. Epub ahead of print 2 July 2022.
- Croucher NJ, Finkelstein JA, Pelton SI, Mitchell PK, Lee GM, et al. Population genomics of post-vaccine changes in pneumococcal epidemiology. *Nat Genet* 2013;45:656–663.
- Derelle R, von Wachsmann J, Mäklin T, Hellewell J, Russell T, et al. Seamless, rapid, and accurate analyses of outbreak genomic data using split *k*-mer analysis. *Genome Res* 2024;34:1661–1673.
- Mostowy R, Croucher NJ, Andam CP, Corander J, Hanage WP, et al. Efficient inference of recent and ancestral recombination within bacterial populations. *Mol Biol Evol* 2017;34:1167–1182.
- Derelle R, Lees J, Phelan J, Lalvani A, Arinaminpathy N, et al. fastlin: an ultra-fast program for Mycobacterium tuberculosis complex lineage typing. *Bioinformatics* 2023;39. Epub ahead of print 1 November 2023.
- Rambaut A, Holmes EC, O'Toole Á, Hill V, McCrone JT, et al. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nat Microbiol* 2020;5:1403–1407.
- Simonsen M, Mailund T, Pedersen CNS. Rapid Neighbour-Joining. In: *Algorithms in Bioinformatics*. Springer Berlin Heidelberg, 2008. pp. 113–122.
- Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol* 2015;32:268–274.
- Tonkin-Hill G, Lees JA, Bentley SD, Frost SDW, Corander J. Fast hierarchical Bayesian analysis of population structure. *Nucleic Acids Res* 2019;47:5539–5549. Epub ahead of print 11 May 2019.
- Argimón S, Abudahab K, Goater RJE, Fedosejev A, Bhai J, et al. Microreact: visualizing and sharing data for genomic epidemiology and phylogeography. *Microb Genom* 2016;2:e000093.
- Lees JA, Tonkin-Hill G, Yang Z, Corander J. Mandrake: visualizing microbial population structure by embedding millions of genomes into a low-dimensional representation. *Philos Trans R Soc Lond B Biol Sci* 2022;377:20210237.
- Didelot X, Croucher NJ, Bentley SD, Harris SR, Wilson DJ. Bayesian inference of ancestral dates on bacterial phylogenetic trees. *Nucleic Acids Res* 2018;46:e134. Epub ahead of print 3 September 2018.
- Croucher NJ, Finkelstein JA, Pelton SI, Parkhill J, Bentley SD, et al. Population genomic datasets describing the post-vaccine evolutionary epidemiology of *Streptococcus pneumoniae*. *Sci Data* 2015;2:150058.
- Robinson O, Dylus D, Dessimoz C. Phylo.io: interactive viewing and comparison of large phylogenetic trees on the web. *Mol Biol Evol* 2016;33:2163–2166.
- Enright MC, Spratt BG. Extensive variation in the *ddl* gene of penicillin-resistant *Streptococcus pneumoniae* results from a hitchhiking effect driven by the penicillin-binding protein 2b gene. *Mol Biol Evol* 1999;16:1687–1695.
- Smith MR. Information theoretic generalized Robinson-Foulds metrics for comparing phylogenetic trees. *Bioinformatics* 2020;36:5007–5013.
- Lees JA, Kendall M, Parkhill J, Colijn C, Bentley SD, et al. Evaluation of phylogenetic reconstruction methods using bacterial whole genomes: a simulation based study. *Wellcome Open Res* 2018;3:33.
- Gladstone RA, Lo SW, Goater R, Yeats C, Taylor B, et al. Visualizing variation within Global Pneumococcal Sequence Clusters (GPSCs) and country population snapshots to contextualize pneumococcal isolates. *Microbial Genomics* 2020;6. Epub ahead of print May 2020.
- Higgs C, Sherry NL, Seemann T, Horan K, Walpole H, et al. Optimising genomic approaches for identifying vancomycin-resistant *Enterococcus faecium* transmission in healthcare settings. *Nat Commun* 2022;13:509.
- Maechler F, Weber A, Schwengers O, Schwab F, Denkel L, et al. Split *k*-mer analysis compared to cgMLST and SNP-based core genome analysis for detecting transmission of vancomycin-resistant

- enterococci: results from routine outbreak analyses across different hospitals and hospitals networks in Berlin, Germany. *Microbial Genomics* 2023;9. Epub ahead of print 30 January 2023.
38. Sundermann AJ, Rangachar Srinivasa V, Mills EG, Griffith MP, Evans E, *et al.* Genomic sequencing surveillance of patients colonized with vancomycin-resistant *Enterococcus* (VRE) improves detection of hospital-associated transmission. *medRxiv*. DOI: 10.1101/2024.05.01.24306710
 39. Rath A, Kieninger B, Mirzaliyeva N, Schmid S, Mester P, *et al.* The genome-oriented surveillance of vancomycin-resistant enterococci shows a clear misclassification of nosocomial transmission events. *Clin Microbiol Infect* 2024;30:1086–1088.
 40. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics* 2014;30:2114–2120.
 41. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, *et al.* SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* 2012;19:455–477.
 42. Page AJ, Taylor B, Delaney AJ, Soares J, Seemann T, *et al.* SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. *Microb Genom* 2016;2:e000056.
 43. McHugh MP, Pettigrew KA, Taori S, Evans TJ, Leanord A, *et al.* Consideration of within-patient diversity highlights transmission pathways and antimicrobial resistance gene variability in vancomycin-resistant *Enterococcus faecium*. *J Antimicrob Chemother* 2024;79:656–668.
 44. Hunt M, Lima L, Shen W, Lees J, IqbalZ, *et al.* AllTheBacteria - all bacterial genomes assembled, available and searchable. *bioRxiv* 2024.
 45. Stevens EL, Carleton HA, Beal J, Tillman GE, Lindsey RL, *et al.* The use of whole-genome sequencing by the federal interagency collaboration for genomics for food and feed safety in the United States. *J Food Prot* 2022;85:755–772. Epub ahead of print 8 March 2022.
 46. Campbell F, Cori A, Ferguson N, Jombart T. Bayesian inference of transmission chains using timing of symptoms, pathogen genomes and contact data. *PLoS Comput Biol* 2019;15:e1006930.
 47. Lees JA, Kremer PHC, Manso AS, Croucher NJ, Ferwerda B, *et al.* Large scale genomic analysis shows no evidence for pathogen adaptation between the blood and cerebrospinal fluid niches during bacterial meningitis. *Microb Genom* 2017;3:e000103.
 48. Colquhoun RM, Hall MB, Lima L, Roberts LW, Malone KM, *et al.* Pandora: nucleotide-resolution bacterial pan-genomics with reference graphs. *Genome Biol* 2021;22:267.
 49. Derelle R, Madon K, Hellewell J, Rodríguez-Bouza V, Arinaminpathy N, *et al.* Reference-free variant calling with local graph construction with SKA lo (SKA). *Mol Biol Evol* 2025:msaf077.
 50. Permana B, Harris PNA, Roberts LW, Cuddihy T, Paterson DL, *et al.* HALviz: an interactive dashboard for visualising and integrating healthcare-associated genomic epidemiological data. *Microb Genom* 2024;10:001200.

The Microbiology Society is a membership charity and not-for-profit publisher.

Your submissions to our titles support the community – ensuring that we continue to provide events, grants and professional development for microbiologists at all career stages.

Find out more and submit your article at microbiologyresearch.org