The sections below describe in more detail the content that could be covered on each of the learning topics and sub-topics listed above, with an emphasis on data interpretation and applications, and with references to case studies.

## Interpretation of genomic QC metrics

Genomic quality metrics are computed at different stages of the sequencing and genome analysis pipeline: raw sequence data, read alignment, variant calling and *de novo* assembly (Table 2). QC metrics are applied to both controls and clinical isolates being sequenced in the same sequencing run to make sure the laboratory processes of DNA extraction, library preparation and sequencing reactions yield sequence data of enough quality for downstream applications.

**Thresholds for quality metrics** should be set beforehand. Technical evaluations of laboratory and whole-genome sequence analysis pipelines have defined key quality metrics and set thresholds for these, either as a single value or a range of values.[8] Other studies[9,10] have defined "warning" and "failure" thresholds by selecting a more and less stringent value for metrics exhibiting less and more variation between samples, respectively.

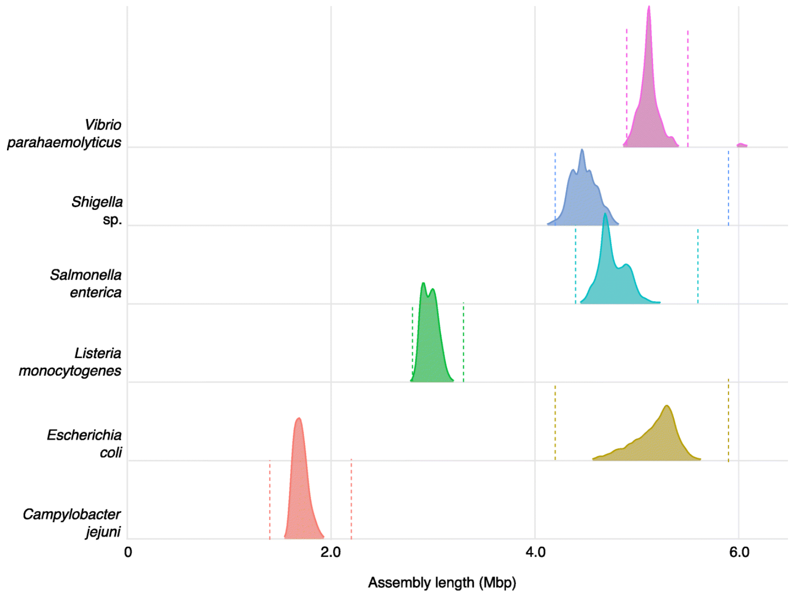


*Figure 1 Distribution of genome assembly lengths for Illumina sequenced isolates of different bacterial organisms*[11]

As these validation studies have shown,[8,10,11] **QC thresholds are often organism specific**, not surprisingly considering that the size and repetitive structure of each microbial genome will determine the value of metrics such as expected assembly length (Figure 1), mean read coverage or number of contigs (Figure 2).

The **whole-genome analysis pipeline used** will also influence the exact values of QC thresholds. Recent studies have assessed the effect of different SNP calling pipelines[12] and key bioinformatic parameters affecting genetic distance calculations.[13]

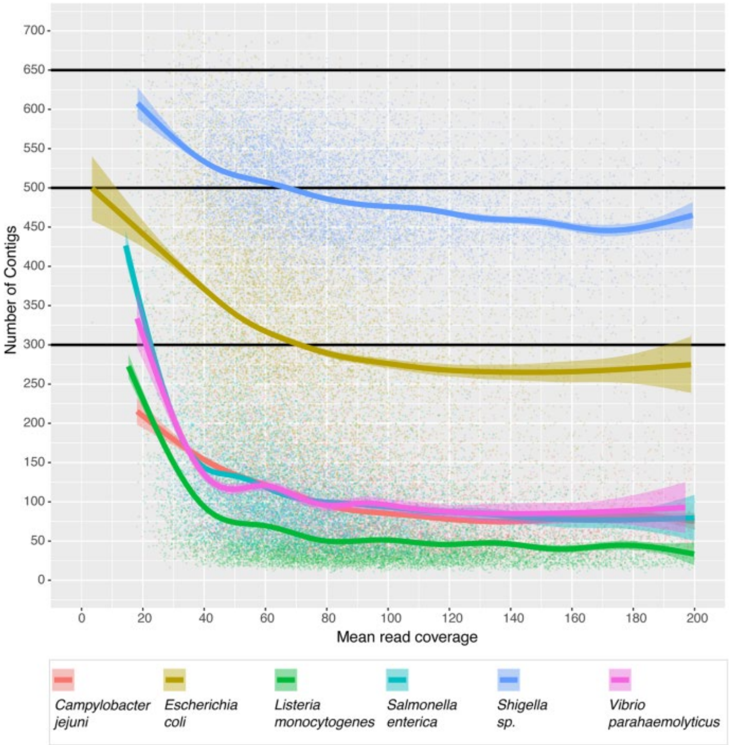


*Figure 2 Mean coverage vs number of contigs for lengths for Illumina sequenced isolates of different bacterial organisms* [11]

In addition to indicating the quality of individual isolate genomes, QC metrics are also used to measure the performance of each end-to-end sequencing process in terms accuracy, precision, reproducibility, and repeatability.[1410] Repeatability is a measure of within-run precision while reproducibility a measure of between-run precision.

*Table 2 Evaluated performance metrics and their corresponding definitions and formulas* [9]

| Metric | Definition | Formula | Assay-specific definitions |
|---|---|---|---|
| Accuracy | The likelihood that results of the assay are correct | Accuracy = 100% × (TP + TN)/(TN + FN + TP + FP) | True-positive result (TP) |
| Precision | The likelihood that detected results of the assay are truly present | Precision = 100% × TP/(TP + FP) | False-negative result (FN) |
| Repeatability | Agreement of the assay based on intra-assay replicates | Repeatability = 100% × (no. of intra-assay replicates in agreement)/(total no. of intra-assay replicates) | Intra-assay replicate |

| | | | |
|---|---|---|---|
| Reproducibility | Agreement of the assay based on inter-assay replicates | Reproducibility = 100% × (no. inter-assay replicates in agreement)/(total no. inter-assay replicates) | Inter-assay replicate |

The QC metrics of positive and negative control sequenced along with clinical strains in the same sequencing run is also encouraged to assess the quality of individual runs. A recent study[8] included a clinical strain of the target sequenced organism (MRSA) as a positive control to control for the accuracy of base calling by the sequencer; a strain of a non-target organism (*Escherichia coli* NCTC12241) as a negative control to ensure lack of cross-contamination and absence of target genetic markers (*S. aureus* sequence type and *mec* gene); and a no-template (water) control (Figure 3). More importantly, the positive control was used to control for base calling accuracy by the sequencer, and used multiple repeat sequences of the control to define a permitted range of SNPs different to the mapping reference for this control (equating to 3 standard deviations from the mean).

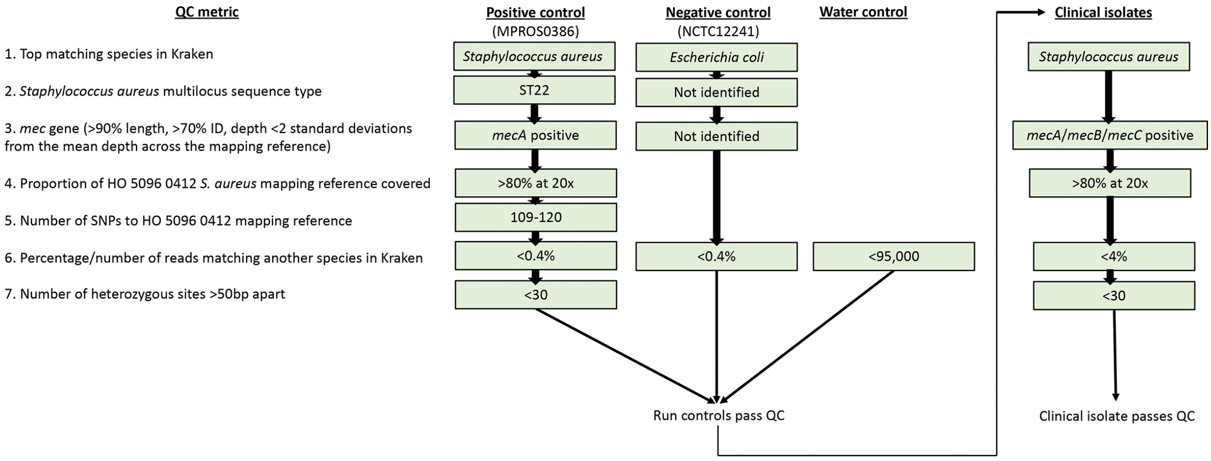

QC metric
1. Top matching species in Kraken
2. *Staphylococcus aureus* multilocus sequence type
3. *mec* gene (>90% length, >70% ID, depth <2 standard deviations from the mean depth across the mapping reference)
4. Proportion of HO 5096 0412 *S. aureus* mapping reference covered
5. Number of SNPs to HO 5096 0412 mapping reference
6. Percentage/number of reads matching another species in Kraken
7. Number of heterozygous sites >50bp apart
Positive control
(MPROS0386)
*Staphylococcus aureus*
ST22
*mecA* positive
>80% at 20x
109-120
<0.4%
<30
Negative control
(NCTC12241)
*Escherichia coli*
Not identified
Not identified
<0.4%
Water control
<95,000
Clinical isolates
*Staphylococcus aureus*
*mecA/mecB/mecC* positive
>80% at 20x
<4%
<30
Run controls pass QC
Clinical isolate passes QC

*Figure 3 QC flowchart for passing/failing controls and clinical isolates during clinical MRSA sequencing* [8]

**Detecting contamination** with different strains of the same species or species different from the target organism is another important purpose of QC of genomic data. Sequencing directly from clinical plates, even when targeting individual colonies for sequencing, runs an increased risk of contamination as more than one strain or a growing plate contaminant could be inadvertently sequenced. Introduction of contaminants can occur at many stages in the generation of bacterial sequence data. For example, cross-contamination can also occur during preparation of genomic DNA or sequencing-libraries.[15]

**Heterozygous sites** detected after read mapping and variant calling are an important metric of same-species contamination when sequencing haploid microbial genomes. Sequencing a mixture of different strains of the same haploid organism will result in the chromosome sites differing between strains being called as heterozygous alleles instead of homozygous alleles. A recent study[8] observed that the majority of heterozygous SNPs in sequenced clinical strains of MRSA clustered together within particular 'hot-spot' locations, putatively attributable to homolog/repetitive sequences. In the pure isolates (0% contamination) sequenced, the majority of heterozygous SNPs were found to be <50 bp apart and therefore only heterozygous SNPs more than >50 bp apart were considered when estimating the proportion of same-species contamination. A cut-off of non-clustered heterozygous SNPs was also established to differentiate contaminated runs (Figure 3).

Not accounting heterozygous SNPs can lead to erroneously high relatedness in the event that bases at heterozygous positions are excluded when calculating pairwise SNP distances, which is standard practice. It has also been showed that within-species contamination causes errors that confound clustering analyses, while between-species contamination generally does not.[16]

Different-species contamination can be detected using taxonomic classifiers like Kraken[17], originally designed for metagenomic studies; or tools like ConFinder[18] which are based in the detection of alleles in ribosomal MLST genes. It has been shown that different-species contaminant DNA is a major source of false genetic variability in bacterial sequencing experiments.[19]

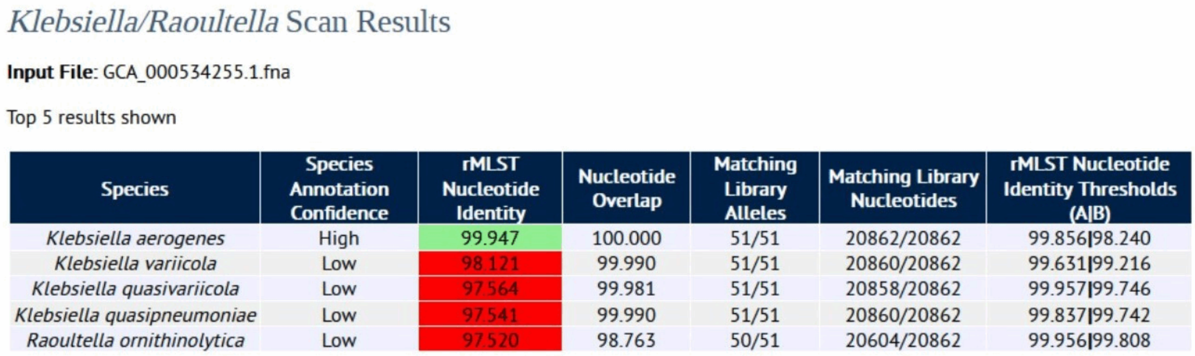
## Interpretation of speciation and strain typing results

Knowing the identity of the pathogen causing any infection allows clinicians to identify appropriate treatment (e.g. to give antibiotics for a bacterial respiratory infection but avoid giving them for a viral infection) and to determine any infection control measures that may be required to prevent its spread.

Unbiased whole-genome sequencing directly from the patient's clinical specimen has been proposed to detect the cause of an infection (diagnostic metagenomics). The field of diagnostic metagenomics is still emerging. The results of proof-of-concept studies[20] using Oxford nanopore sequencing are promising.[21–23]

If identification of the infection-causing pathogens can be made directly from the culture using phenotypic methods, there is little obvious additional clinical utility in sequencing the genome of the causative agent. Nonetheless, when target pathogens are sequenced for other applications (e.g. outbreak investigation, detection of AMR), speciation from genomic data is used to confirm the **identity of the target organism** and inform QC.

A quick and accurate method to confirm the identify of bacterial species is ribosomal MLST nucleotide identity (rMLST-NI).[24] Ribosomal MLST (rMLST) is a universal, bacterial domain-wide approach that indexes the protein-encoding genes of the ribosome and has been shown to reconstruct phylogenetic and taxonomic groups accurately. rMLST profiles are defined based on the numeric allelic indices of the 53 rMLST loci. The rMLST nucleotide identity of the sequenced genome can be calculated against the rMLST profiles of bacterial species available to date to identify the closest match and determine their species, similar to how sequence types (STs) are assigned using MLST loci. rMLST-NI results were found to agree with those obtained by whole-genome average nucleotide identity methods (OrthoANIu and FastANI).[24]



*Klebsiella/Raoultella* Scan Results

**Input File**: GCA_000534255.1.fna

Top 5 results shown

| Species | Species Annotation Confidence | rMLST Nucleotide Identity | Nucleotide Overlap | Matching Library Alleles | Matching Library Nucleotides | rMLST Nucleotide Identity Thresholds (A\|B) |
|---|---|---|---|---|---|---|
| *Klebsiella aerogenes* | High | 🟩 99.947 | 100.000 | 51/51 | 20862/20862 | 99.856\|98.240 |
| *Klebsiella variicola* | Low | 🟥 98.121 | 99.990 | 51/51 | 20860/20862 | 99.631\|99.216 |
| *Klebsiella quasivariicola* | Low | 🟥 97.564 | 99.981 | 51/51 | 20858/20862 | 99.957\|99.746 |
| *Klebsiella quasipneumoniae* | Low | 🟥 97.541 | 99.990 | 51/51 | 20860/20862 | 99.837\|99.742 |
| *Raoultella ornithinolytica* | Low | 🟥 97.520 | 98.763 | 50/51 | 20604/20862 | 99.956\|99.808 |

*Figure 4 rMLST nucleotide identity webserver results for scanning UCI 27 (NCBI Assembly entry GCA_000534255.1, Klebsiella aerogenes)*

Taxonomic classifiers based on 16S rRNA-based species identification and Kraken, the latter based on scanning kmers of the sequenced genome against databases of complete microbial genomes, are also commonly used. Another approach consists in the detection of specific genetic markers (genes, deletions, SNPs) that are known to be specific of the target pathogen. For example, SNPs specific to all subspecies of the *Mycobacterium tuberculosis* complex (MTBC) have been identified and used for typing purposes.[25]

At the strain level, high-resolution strain typing methods based on core-genome and whole-genome MLST have been developed for multiple bacterial species. For clonal pathogens like *Mycobacterium tuberculosis* complex[26,27] (https://github.com/jodyphelan/TBProfiler) and *Salmonella enterica* serovar Typhi[28] (https://github.com/katholt/genotyphi , genotyping schemes based on the detection of lineage and sub-lineage specific SNPs have also been developed. The detection of SNPs specific of multiple lineages or sub-lineages can also point to identify cases of mixed infections.

# How to interpret phylogenetic trees in the context of genomic epidemiology

## What are phylogenetic trees and how are they reconstructed?

Before teaching how phylogenetic trees are interpret for infectious diseases (ID) epidemiology, it is important that the basic concepts of phylogenetics are introduced, including nomenclature and assumptions of phylogenetic tree reconstructions.

A phylogenetic tree depicts estimated evolutionary relationships between taxa - these can be species, strains or even genes. In the context of ID epidemiology, phylogenetic trees are commonly used to define evolutionary relationships between clinical strains of the same microbial species.

Phylogenetic trees are reconstructed based on the **assumption that microbes reproduce clonally** (although this assumption does not always holds true). During clonal reproduction, microbial progenitor cells replicate their DNA at high fidelity. Despite this, random errors in DNA replication may still occur, resulting in a clonal progeny that will inherit these genetic replication ‘errors’ (i.e. mutations) in their DNA and may not be strictly identical to their progenitor cells. Microbial strains that have recently originated from the same progenitor cell are thus expected to share identical genomes, or have diverged at most by only a few genetic differences. The number and pattern of shared mutations between strains can be used to reconstruct their genealogical and evolutionary relationships.

Next, it’s important to introduce **phylogenetic nomenclature**, as terms like “clade”, “tips”, “topology” or “branches” are commonly used in the field of ID genomic. In short: isolated microbial strains are depicted on the tips (or leaves) of the tree (i.e. taxa), whereas the internal nodes of the tree denote their hypothetical ancestors. Nodes and taxa are connected by branches, the length of which represent genetic distances between connected groups. Groups of strains (taxa) that share the same common ancestor form a monophyletic group (also known as clade). A group of strains that descends from a common ancestor, but does not include all descendants, is called paraphyletic.

There are multiple **online resources on how to read phylogenetic trees** that introduce these phylogenetic concepts and nomenclature including. The EBI course on phylogenetics, for example, places an emphasis on how to read and interpret phylogenetic trees (https://www.ebi.ac.uk/training/online/courses/introduction-to-phylogenetics/). The US CDC course module “How to read a phylogenetic tree”, describes the anatomy of phylogenetic trees and how to interpret them in the context of transmission (https://www.cdc.gov/amd/training/covid-toolkit/module1-3.html).

## How are phylogenetic trees interpreted in ID epidemiological investigations?

**Appendix 1** of this document includes an **example of exercises on how to interpret phylogenetic trees**, with a particular focus on extracting strain relatedness information. Reading phylogenetic trees correctly may be relatively straightforward for an expert user, but should not be taken for granted.

A powerful approach to teach learners these concepts would be to take them through the variety of **proof-of-concept and case studies** that applied genomic epidemiology and phylogenetic trees to investigate microbial transmission. There are multiple key studies captured by governmental agencies, reports and literature reviews that can be used to develop teaching materials on this topic.

The U.S. Food & Drug Administration (FDA) federal agency coordinates a network of laboratories (**GenomeTrakr network**) that make use of whole-genome sequencing for pathogen identification and outbreak detection of foodborne pathogens (https://www.fda.gov/food/whole-genome-sequencing-wgs-program/examples-how-fda-has-used-whole-genome-sequencing-foodborne-pathogens-regulatory-purposes). This website includes multiple case studies on the use of genomic epidemiology to identify and understand the source of **foodborne outbreaks**:

- to determine which illnesses are part of an outbreak and which are not;
- to determine which ingredient in a multi-ingredient food is responsible for an outbreak;
- to identify geographic regions from which a contaminated ingredient may have originated;
- to differentiate sources of contamination, even within the same outbreak;
- to link illnesses to a processing facility even before the food product vector has been identified;
- to link small numbers of illnesses that otherwise might not have been identified as common outbreak;
- and to identify unlikely routes of contamination.

As an example, Figure 5 includes 35 genomes of *Salmonella enterica* collected as part of an investigation of a major **salmonellosis outbreak** associated with a meat processing facility in New England, US. The phylogenetic analysis of WGS data, from *S. enterica* strains isolated from ingredient suppliers, patients who consumed finished products, and historically and geographically disparate food sources, revealed a recent and common origin for outbreak strains from the implicated food facility (clades E and F in Figure 5). Case studies like this one can be used to exemplify the use of phylogenetics in the investigation of food-borne outbreaks, the expected clustering in tight phylogenetic clades of outbreak cases, and the conclusions that can be derived from.

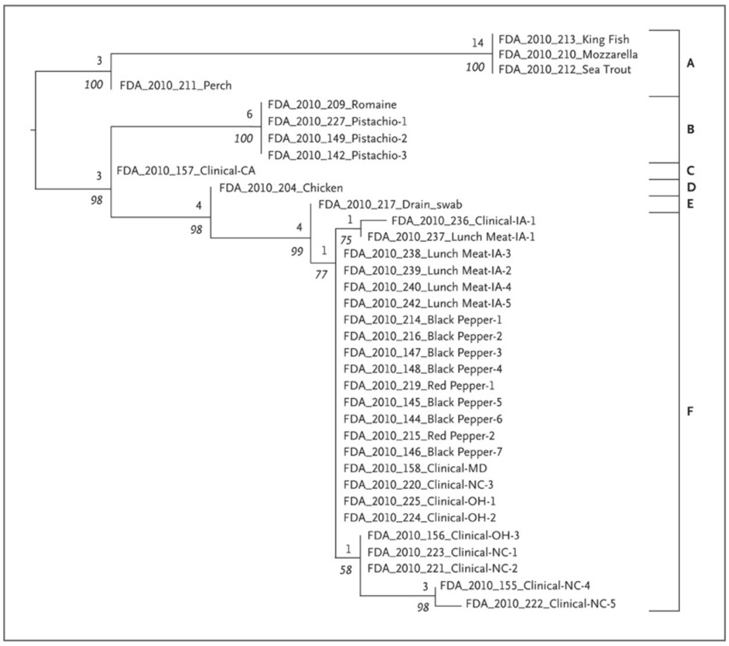

3
100
FDA_2010_211_Perch
14
100
FDA_2010_213_King Fish
FDA_2010_210_Mozzarella
FDA_2010_212_Sea Trout
A
6
100
FDA_2010_209_Romaine
FDA_2010_227_Pistachio-1
FDA_2010_149_Pistachio-2
FDA_2010_142_Pistachio-3
B
3
98
FDA_2010_157_Clinical-CA
C
4
98
FDA_2010_204_Chicken
D
4
99
FDA_2010_217_Drain_swab
E
1
77
1
75
FDA_2010_236_Clinical-IA-1
FDA_2010_237_Lunch Meat-IA-1
FDA_2010_238_Lunch Meat-IA-3
FDA_2010_239_Lunch Meat-IA-2
FDA_2010_240_Lunch Meat-IA-4
FDA_2010_242_Lunch Meat-IA-5
FDA_2010_214_Black Pepper-1
FDA_2010_216_Black Pepper-2
FDA_2010_147_Black Pepper-3
FDA_2010_148_Black Pepper-4
FDA_2010_219_Red Pepper-1
FDA_2010_145_Black Pepper-5
FDA_2010_144_Black Pepper-6
FDA_2010_215_Red Pepper-2
FDA_2010_146_Black Pepper-7
FDA_2010_158_Clinical-MD
FDA_2010_220_Clinical-NC-3
FDA_2010_225_Clinical-OH-1
FDA_2010_224_Clinical-OH-2
1
58
FDA_2010_156_Clinical-OH-3
FDA_2010_223_Clinical-NC-1
FDA_2010_221_Clinical-NC-2
3
98
FDA_2010_155_Clinical-NC-4
FDA_2010_222_Clinical-NC-5
F

*Figure 5* Outbreak Strains of *Salmonella enterica* Subtype Montevideo

Other case studies on the use of phylogenetics to investigate foodborne bacterial outbreaks can be extracted from this review.[29] For example, an investigation into the source of *Listeria monocytogenes* contaminating products in an ice cream firm (Producer) confirmed that the *L. monocytogenes* strain arose from a reservoir population in a supplier's facility, rather than in the Producer's facility. This type of **source attribution** investigations show how the phylogenetic tree should be interpreted: the genetic diversity of the contaminating strain (i.e. source) encloses within the same phylogenetic clade the diversity of the strain's contaminated product.

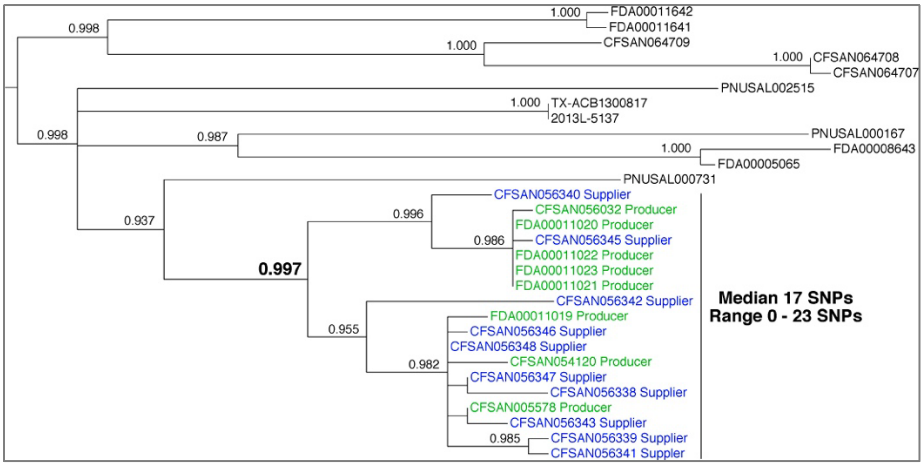


*Figure 6* Phylogenetic analysis of genome sequences obtained from *Listeria monocytogenes* isolated from 2016 ice cream samples and the environment of a supplier.

Public Health England's 2018 report on "Implementing pathogen genomics" (https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/731057/implementing_pathogen_genomics_a_case_study.pdf) includes another three case studies that can be used.

There are also multiple early proof-of-concept and recent studies on the application of phylogenetics to investigate **hospital outbreaks**.

For example, an early proof-of-concept study demonstrated that sequencing methicillin-resistant *Staphylococcus aureus* (MRSA) isolates from a suspected outbreak in a special care baby unit (SCBU) could link previously unsuspected cases (with no apparent epidemiological links and different antibiograms) to the same outbreak, **identify the source of the outbreak** (a health-care worker) and inform the infection control actions that brought the outbreak to a close. From an **interpretation point of view** it is important to note that the MRSA genomes of the health-care staff (i.e. the source of the outbreak) formed a **cloud of diversity** in the phylogenetic tree that enclosed the rest of outbreak isolates from infants, with **close genetic matches** (as shown by short branches and small SNP distances) to isolates from infants, and with two colonies **clustering closer towards the root** of the tree (Figure 7). All these phylogenetic observations are a clear indication of the health-care staff being the **source of the outbreak**.

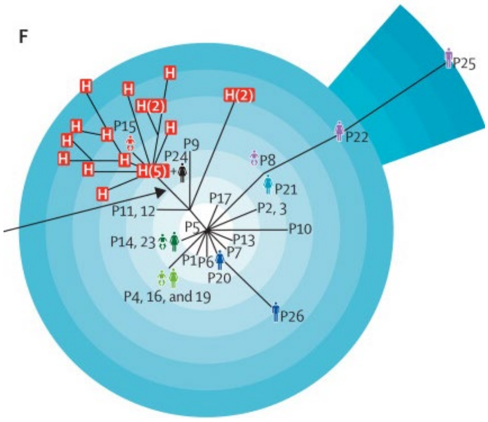


Figure 7 Phylogeny of the MRSA SCBU outbreak

WGS has also been used for **source tracking in hospital settings** to demonstrate the detection of transmission events from hospital water. In a burns care ward and critical care ward in the UK, investigators found that *Pseudomonas aeruginosa* infecting burns patients clearly originated from hospital water sources.[30] In the phylogenetic tree, two of the three patients investigated (patient 1 and 4) fell within the cluster originating from shower water, indicating that shower hydrotherapy was the most likely source of infection.

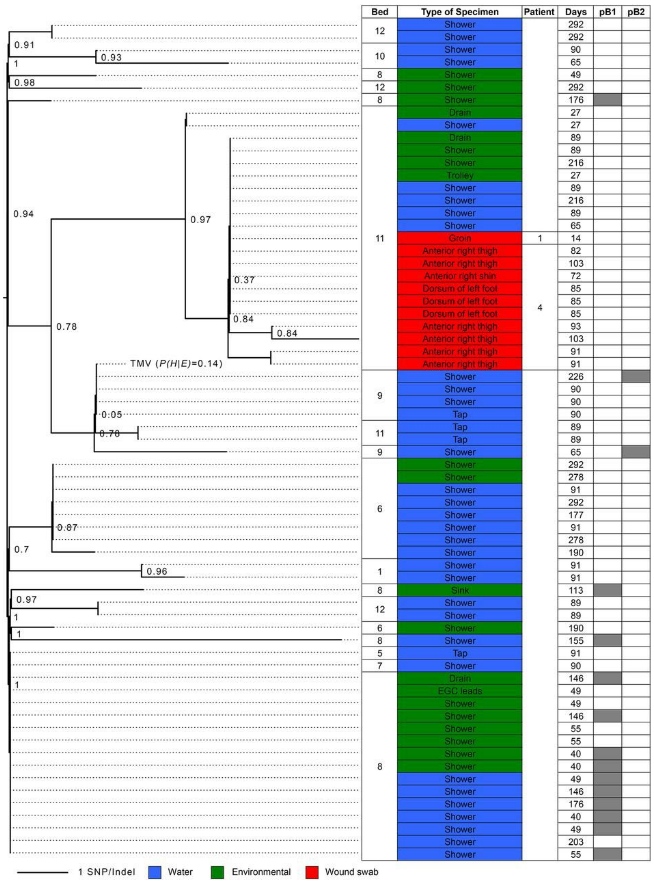


Figure 8 Phylogenetic tree of *P. aeruginosa* clinical and environmental isolates from [30]

Genomic epidemiology has also been applied to study the **transmission of pathogens in the community**, being *Mycobacterium tuberculosis* one of the best exemplar pathogens. Genomic epidemiology studies have provided valuable insights into the phylo-geography of *Mycobacterium tuberculosis* complex (MTBC), its evolutionary pathways and population and nosocomial transmission helping to distinguish between reinfection and re-activation and detect laboratory cross-contamination. Contact tracing complemented with MTBC genotyping is considered an important means of understanding person-to-person transmission. Multiple literature reviews[31–33] summarise case studies on the use of genomic epidemiology and phylogenetics applied to the study of *Mycobacterium tuberculosis*.

## Visualisation of genomic and epidemiological data

**Visualisation and annotation of phylogenetic trees** with epidemiological data (e.g. patient identifier, collection time, location, etc.) are a simple and powerful approach to intuitively and visually investigate outbreaks.

Web-based tools like **iTOL** (https://itol.embl.de/) can be used to visualise phylogenetic trees along with annotations of epidemiological data, as shown in Figure 9 , taken from a study investigating an outbreak of multidrug-resistant tuberculosis (MDR-TB) in Denmark.[34] Basic epidemiological metadata (such as host ids, region of birth, and drug susceptibility labels) annotated on top of the topological information of the tree, allows to visually identify and confirm suspected outbreaks.

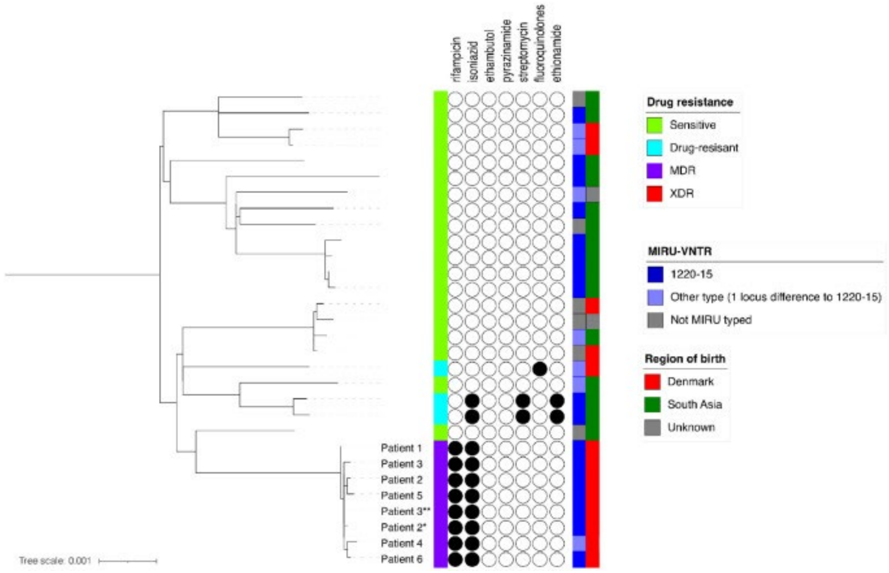


*Figure 9 Phylogenetic tree and metadata of an MDR-TB outbreak strain in Denmark and contextual strains plotted using iTol*

More specialised tools such as **Microreact** (https://microreact.org/) or **Nextstrain** (https://nextstrain.org/) include purpose-built functionality for genomic epidemiology investigations.

Beyond local outbreak investigations, Microreact can be used to visualise the temporal and geographical distribution and evolution of entire pathogen populations. A recent example of that is the use of Microreact to investigate the **epidemiology of SARS-CoV-2 in the UK** using viral genome sequences. Labelling viral variants on the phylogenetic tree, maps and timelines allow to detect what variants are largely responsible for the rise in cases at a particular time.

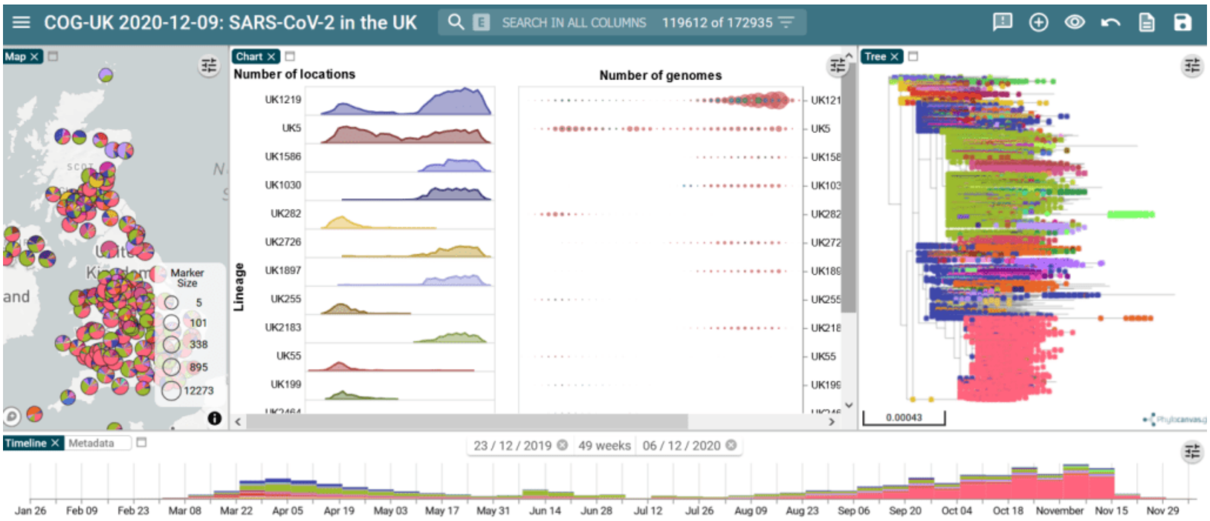


*Figure 10 Visualisation of SARS-CoV-2 genome data in the UK from Microreact*

In the **investigation of hospital outbreaks**, the use of **timeline plots** showing patients' stays in individual hospital wards and departments, along with phylogenetic information, has been commonly used to visually identify the source, in terms of hospital locations or host spreaders, of nosocomial outbreaks.

As an example, the phylogenetic trees and timeline plots of the MRSA clones detected in two different hospital intensive care units (ICUs) in Thailand allowed the investigators to identify one patient on each ICU (T12 and T126, Figure 11) as the source of most transmission events (super-spreader).

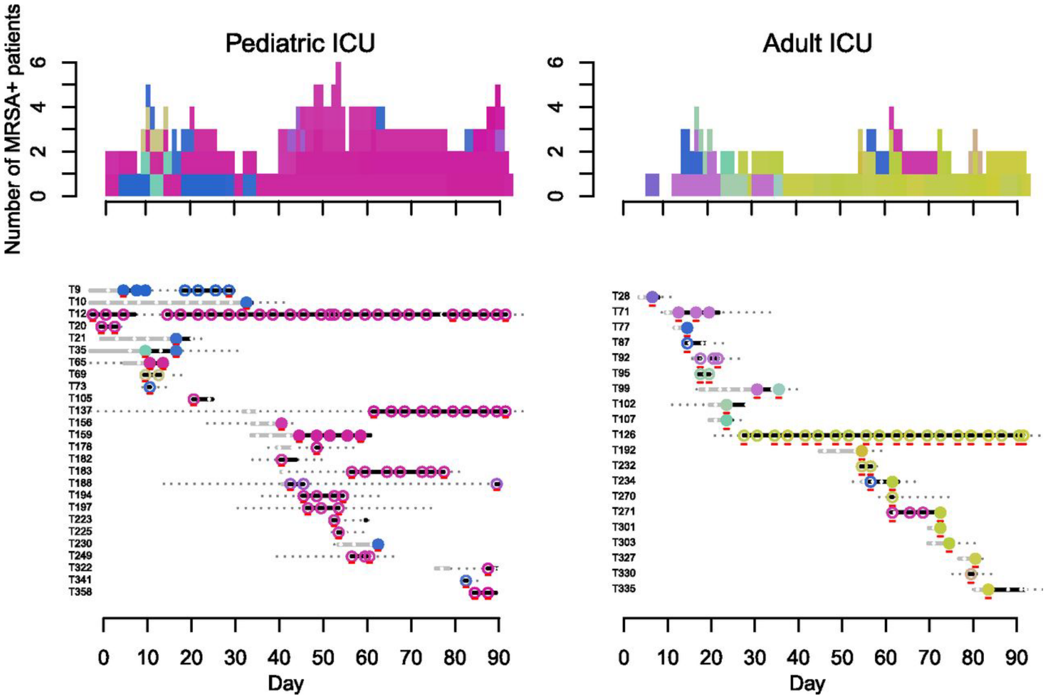


Figure 11 Dynamics of MRSA clones on a paediatric (left) and adult (right) ICUs in Thailand

A timeline of spatiotemporal movements and overlaps of patients for specific transmission clusters (outbreak clones) can help identify the hospital wards where the outbreak originated and how it may have spread to other wards. In this type of plots, each row represents the hospital admission period(s), coloured boxes patient visits to hospital wards (or rooms, departments, etc.) and symbols (e.g. circles) sample collection dates. Figure 12 shows an exemplar of transmission cluster of an *Enterococcus faecium* clone spanning two hematology wards and involving 7 patients. Strong genetic (SNP distances below 6 SNPs) and epidemiological links (stays in the same ward at the same time) point to transmission of this clone in room A3 (coloured in orange) among four patients (C015, C023, C009 and D021), followed by spread of this clone in different rooms of ward B among another four patients (D021, D022, D010 and D045).

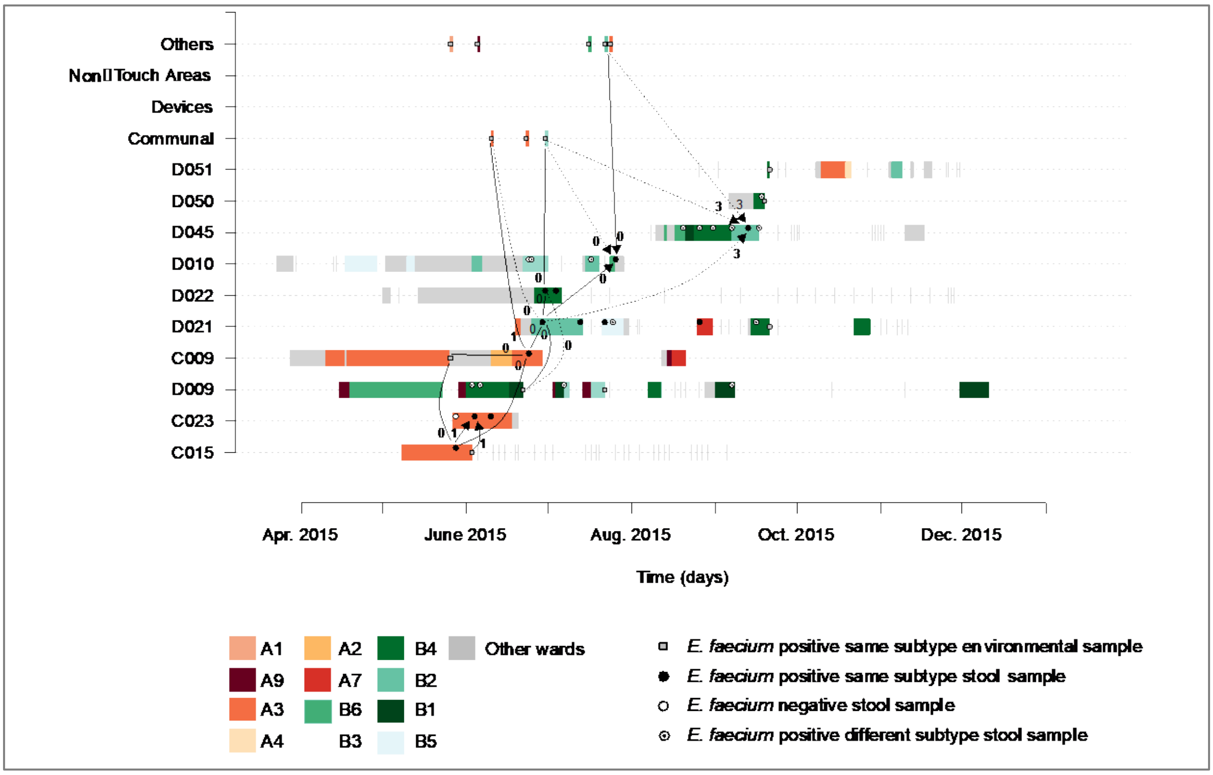


Figure 12 Exemplar of nosocomial transmission of an *E. faecium* clone in two hematology wards

## How genetic relatedness thresholds are applied and interpreted

The detection of pathogen transmission is informed by the determination and interpretation of **genetic distances** from microbial genomic data. The simplest way to establish genetic relatedness between microbial strains is to count the number of nucleotide differences (i.e. the number of single-nucleotide polymorphisms [SNPs]) between their whole or core-genome sequences. The SNP cut-off approach places two individuals in the same putative transmission cluster (i.e. outbreak) if the genetic relatedness of their microbial isolates is below a pre-defined number of SNPs.

Common approaches to determine SNP cut-offs are based on the maximum within-host diversity observed (assuming that that's the maximum pre-existing diversity that could be transmitted from the source to a recipient) or the distribution of genetic distances between strains from cases with confirmed epidemiological links.

Genetic relatedness thresholds have been proposed above which recent microbial transmission (ideally defined within a specific time frame) can be ruled out, while distances below indicate probable transmission. It is increasingly acknowledged that epidemiological follow-up (i.e. detection of common epidemiological links) is needed to confirm definite transmission.

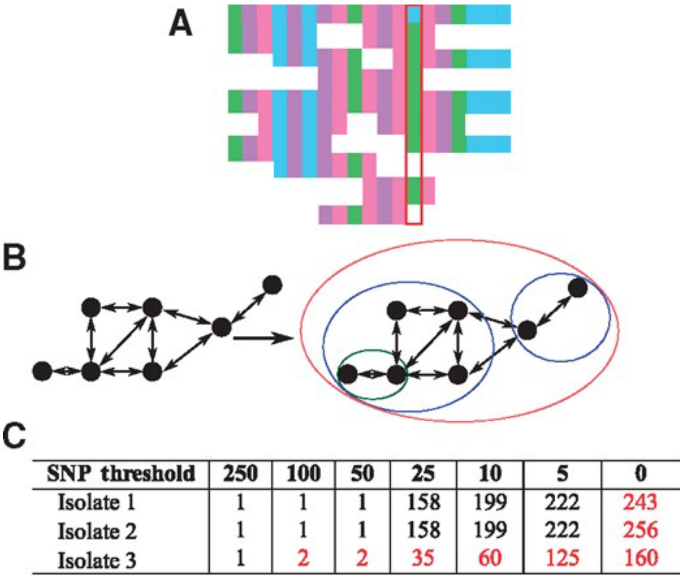
SNP cut-offs are a simple and intuitive measure of genetic relatedness that can be interpreted by non-expert users in a clinical setting. **Limitations of the SNP cut-off approach** include that the likelihood of direct transmission below the cut-off cannot be inferred, and the directionally (i.e. who infection whom) cannot be determined either. Others warn that differences is local epidemiology, e.g. the presence of recently expanded and predominant circulating clones, can lead to the majority of cases being genetically linked below a SNP cut-off. In any case, the **identification of common epidemiological links** (such as visits to the same hospital ward, unit, or clinic, shared residential postcodes, or management by the same health-care worker) are still essential to confirm definite transmission, identify the place/host of transmission and direct infection control interventions.

SNP cut-offs have been reported for MRSA[35], *M. tuberculosis*[31] and *K. pneumoniae*[36], and foodborne pathogens[29] (not an exhaustive list).

For *M. tuberculosis*, studies conducted in 2013 to 2017 demonstrated that genetic distances between strains from patients with confirmed epidemiological links were generally within the range of 0 to 5 SNPs. Based on the mutation rate and epidemiological observations, a cut-off value of fewer than six SNPs has been proposed to indicate recent transmission, defined as transmission that has occurred in the last 3 years; while strains being >12 SNPs apart were not considered involved in direct transmission.[31]

Similarly related to SNP cut-offs, core-genome and whole-genome MLST schemes can also be used to detect microbial transmission, in addition to their use as typing schemes, provided that the number of allele differences compatible with probable transmission are known (effectively a SNP cut-off too).

Hierarchical SNP relatedness approaches (effectively the application of SNP cut-offs at different relatedness levels) has also been described. *SnapperDB* was developed by Public Health England (PHE, now UKHSA) to quantify SNP relatedness and derive an isolate level nomenclature termed the "SNP Address". This applies multi-threshold SNP typing to describe an isolate's relatedness (allele profile, SNP address) in relation to a population of previously typed strains. Clustering is performed at seven descending thresholds of SNP distance; 250, 100, 50, 25, 10, 5, and 0. This clustering results in a seven-digit code where each number represents the cluster membership at each descending SNP distance threshold. Instead of applying a single SNP cut-off to rule in or out probable transmission, this approach allows for a directed and hierarchical investigation of genetically related cases.



| SNP threshold | 250 | 100 | 50 | 25 | 10 | 5 | 0 |
|---|---|---|---|---|---|---|---|
| Isolate 1 | 1 | 1 | 1 | 158 | 199 | 222 | 243 |
| Isolate 2 | 1 | 1 | 1 | 158 | 199 | 222 | 256 |
| Isolate 3 | 1 | 2 | 2 | 35 | 60 | 125 | 160 |

*Figure 13*

# How to interpret genotypic predictions of antibiotic resistance from microbial genomes

## Early proof-of-concept studies

Antibiotic resistance in bacteria is mediated by the acquisition of new genes or genetic variants in existing regions of the core or accessory genome that enable the organism to avoid the toxic effects of the drug. Multiple proof-of-concept studies demonstrated that, in principle, it is possible to use WGS to detect genes and genomic variants that are known to cause antibiotic resistance.

For *Staphylococcus aureus,*[1] an early study showed that genotype-based antibiotic-resistance prediction was comparable to that obtained by gold-standard phenotypic methods, with a sensitivity and specificity of 99.1% and 99.6%, respectively, across 12 antibiotics.

For *Mycobacterium tuberculosis*, the application of WGS to determine drug resistance has been particularly attractive, given the slow-growing rate of this bacterium and long turnaround times of phenotypic methods to yield susceptibility results. In an early proof-of-concept study,[2] the authors used WGS to investigate the case of a patient with extensively drug-resistant (XDR) tuberculosis. The reference laboratory had reported resistance to nine antibiotics, and the authors detected mutations that were consistent with resistance to these nine drugs. They concluded that whole-genome sequencing had the potential to diagnose drug resistance in *M. tuberculosis* within weeks to days.

A retrospective WGS study on *Escherichia coli* and *Klebsiella pneumoniae* isolates demonstrated that WGS was as sensitive and as specific as currently used phenotypic methods at predicting antimicrobial sensitivity.[3]

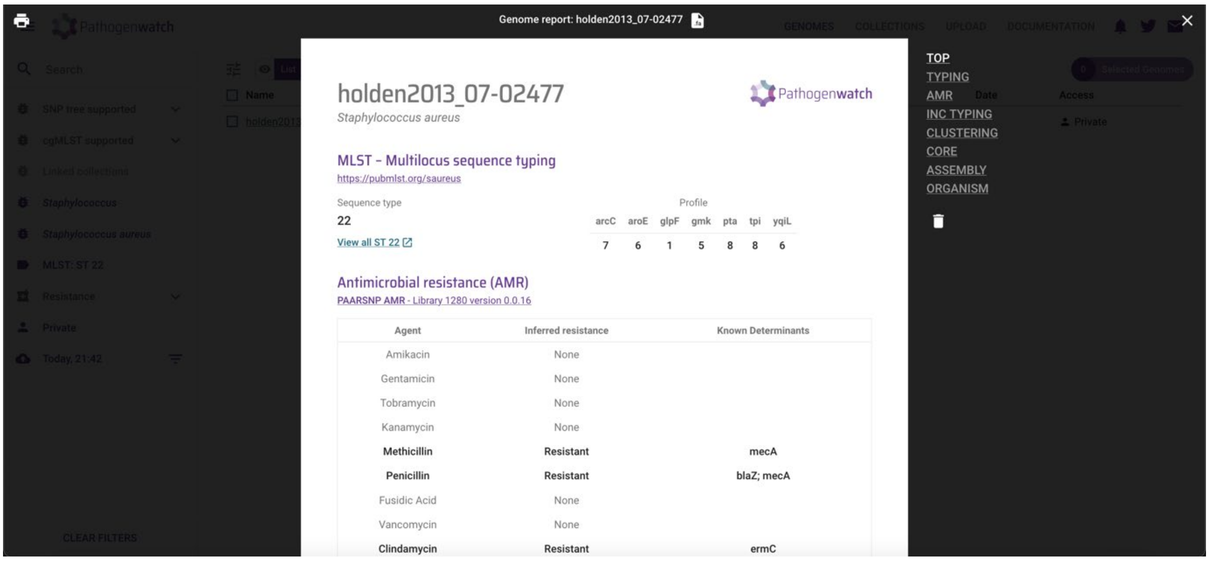
## Available approaches, databases and tools

The most common approach to predict antibiotic resistance from genome sequences is the look-up table or rule-based approach, wherein the genome is scanned for the presence of genetic markers encoded in a database of antibiotic resistance (ABR) determinants. The absence of any known antibiotic resistance determinant is interpreted as susceptibility to that specific antibiotic. ABR determinants are genetic markers such as single acquired genes, multiple acquired genes (e.g. operons), individual mutations or multiple mutations in the same or different genes. The latter two include amino acid changes in protein-coding genes, nucleotide changes in RNAcoding genes or promoter mutations. The creation and maintenance of these databases require an continuous expert curation.

This approach requires a comprehensive knowledge of the genetic basis of resistance for each target antibiotic in the target microbial organism. This is generally not the case for recently licensed or last-line antibiotics, for which not all resistance mechanisms are characterised. It is increasingly acknowledged that ABR genetic determinants must be applied to determine resistance to individual antibiotics (as opposed to entire antibiotic classes) and for individual microbial species.

Machine learning techniques have also been explored to predict phenotypic resistance from whole-genome sequence data,[37,38] although they seem to perform on par with look-up table approaches, while not always providing an intuitive interpretation of ABR predictions. Methods to infer ABR phenotypes from bacterial genomes by closest genomic match, that is, by identifying its closest relatives in a database of genomes with ABR metadata, have also been proposed.[39]

Tools like AMRFinder,[4] CARD Resistance Gene Identifier (RGI),[5] ResFinder,[6] or Pathogenwatch (https://pathogen.watch/) are among the most commonly used bioinformatic tools to determine ABR, which also host underlaying curated databases of ABR genetic markers needed to make these predictions. **Pathogenwatch** is one of most intuitive an easy-to-use web-based platforms for the analysis of bacterial genomes, developed by The Centre for Genomic Pathogen Surveillance (CGPS), UK, that can be used to detect AMR in the genomes of many microbial pathogens. Pre-generated genome assemblies can be directly uploaded as input to this tool. Once uploaded, Pathogenwatch performs strain identification, multi-locus sequence typing (MLST) and resistance prediction in an automated manner. Recently, the website was upgraded with the option to upload raw sequencing reads, those obtained directly from sequencing machines without further bioinformatic processing.



holden2013_07-02477
*Staphylococcus aureus*

MLST – Multilocus sequence typing
https://pubmlst.org/saureus

Sequence type: 22
View all ST 22

| arcC | aroE | glpF | gmk | pta | tpi | yqiL |
|---|---|---|---|---|---|---|
| 7 | 6 | 1 | 5 | 8 | 8 | 6 |

Antimicrobial resistance (AMR)
PAARSNP AMR - Library 1280 version 0.0.16

| Agent | Inferred resistance | Known Determinants |
|---|---|---|
| Amikacin | None | |
| Gentamicin | None | |
| Tobramycin | None | |
| Kanamycin | None | |
| **Methicillin** | **Resistant** | **mecA** |
| **Penicillin** | **Resistant** | **blaZ; mecA** |
| Fusidic Acid | None | |
| Vancomycin | None | |
| **Clindamycin** | **Resistant** | **ermC** |

*Figure 14 Pathogenwatch genome report of an MRSA strain*

For well-studied organisms, like *M. tuberculosis*, there are organism-specific tools like MTBseq[40], Mykrobe[41] and TB-Profiler.[42]

## Assessing the diagnostic accuracy of genotypic determinations with population-based studies

While many publicly available bioinformatics tools have been developed to determine ABR genotypically from whole-genome sequences, based on different curated ABR databases, the quality and diagnostic accuracy of genotypic predictions compared to phenotypic antibiotic susceptibility testing (AST) results cannot be taken for granted. As stated before, the accuracy of genotypic predictions should be assessed for individual antibiotics and bacterial species.

Large and diverse collections of strains with available AST phenotypes and genome sequences are needed to assess the diagnostic performance of genotypic predictions. For binary determinations of ABR (i.e. resistance or non-susceptibility vs. susceptibility), the following diagnostic classifications are commonly used:

- True positive: phenotypically resistant (or non-susceptible) strain with known resistance-conferring genetic determinant(s) detected in their genome.
- True negative: phenotypically susceptible strain in the absence of any known genetic determinant.
- False positive: phenotypically susceptible strain in the presence of a known genetic determinant.
- False negative: phenotypically resistant (or non-susceptible) strain but not carrying known resistance-conferring genetic determinant(s) in their genome.

The number of strains in each of these four categories are used to calculate metrics of diagnostic accuracy such as sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV). As an example, Figure 15 shows the diagnostic accuracy, in terms of sensitivity and specificity, of WGS to predict phenotypic resistance to first-line drugs in *M. tuberculosis.* In this type of studies, a low sensitivity (i.e. high number of false negatives) should be interpreted as a large proportion of phenotypically resistant strains lacking known ABR genetic markers, in other words, resistance is under-called at the population level. On the other hand, a low specificity (caused by a big number of false positives), is indicative of a large number of phenotypically susceptible strains carrying ABR genetic markers, that is, resistance in over-called at the population level. In conclusion, when deciding which bioinformatic tool or database to use for ABR detection from whole-genome sequences is important to identify studies benchmarking these tools and providing metrics of diagnostic accuracy broken down for specific antibiotics and microbial species.

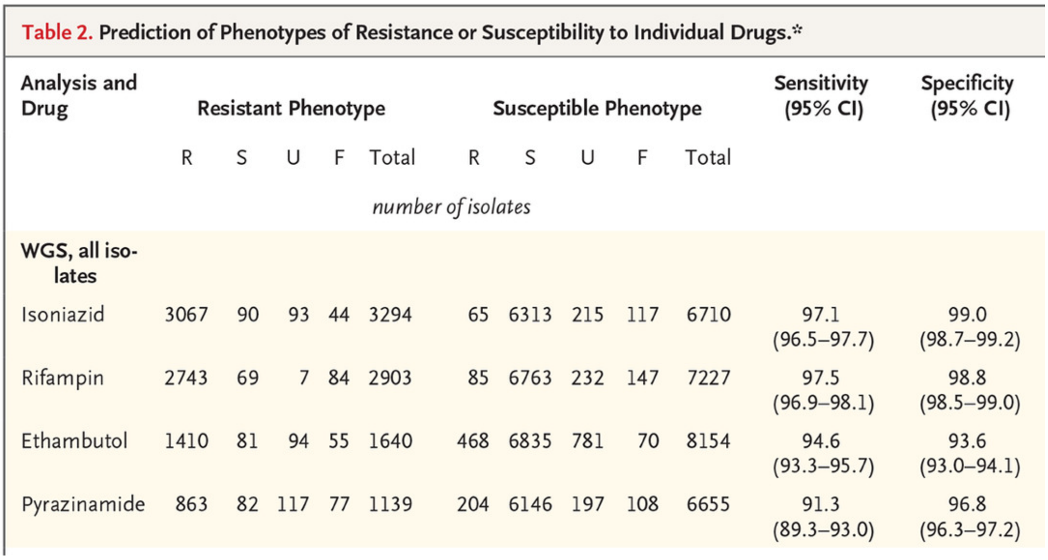


Table 2. Prediction of Phenotypes of Resistance or Susceptibility to Individual Drugs.*

| Analysis and Drug | Resistant Phenotype | | | | | Susceptible Phenotype | | | | | Sensitivity (95% CI) | Specificity (95% CI) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | R | S | U | F | Total | R | S | U | F | Total | | |
| | *number of isolates* | | | | | | | | | | | |
| WGS, all isolates | | | | | | | | | | | | |
| Isoniazid | 3067 | 90 | 93 | 44 | 3294 | 65 | 6313 | 215 | 117 | 6710 | 97.1<br>(96.5–97.7) | 99.0<br>(98.7–99.2) |
| Rifampin | 2743 | 69 | 7 | 84 | 2903 | 85 | 6763 | 232 | 147 | 7227 | 97.5<br>(96.9–98.1) | 98.8<br>(98.5–99.0) |
| Ethambutol | 1410 | 81 | 94 | 55 | 1640 | 468 | 6835 | 781 | 70 | 8154 | 94.6<br>(93.3–95.7) | 93.6<br>(93.0–94.1) |
| Pyrazinamide | 863 | 82 | 117 | 77 | 1139 | 204 | 6146 | 197 | 108 | 6655 | 91.3<br>(89.3–93.0) | 96.8<br>(96.3–97.2) |

*Figure 15 Diagnostic accuracy of WGS to predict phenotypic resistance to first-line drugs in M. tuberculosis*

## Limitations and sources of genotype-phenotype discrepancies

For some pathogens and antimicrobials, the predictive sensitivity and specificity of WGS for inferring AMR phenotypes are still too low for practical application. When comparing the concordance between phenotypic and genotypic AMR it is essential to consider the reasons that errors may occur. Three broad reasons for systematic errors are:[43]

(a) Heteroresistance and an inadequate limit of detection of WGS.
(b) Flaws with phenotypic AST.
(c) Incomplete understanding of the genotypic basis of phenotypic resistance.

Many bacterial species and antibiotic classes exhibit heteroresistance, a phenomenon in which a susceptible bacterial isolate harbours a resistant subpopulation that can grow in the presence of an antibiotic and cause treatment failure.[44] A drug-susceptible (usually wildtype; WT) and drug-resistant (usually mutant) organisms can co-exist in the same clinical specimen which can result from the concurrent presence of two different strains (mixed infection) or from a changing bacterial subpopulation within the same strain (clonal evolution) (Figure 16).[45] The capacity of phenotypic and genotypic assays to detect heteroresistance largely depends on the technique applied. Phenotypic antibiotic-susceptibility testing can determine whether 1% or more of the bacterial population is antibiotic resistant. Accordingly, genotypic assays should be able to detect the simultaneous presence of WT and mutant sub-populations of the relevant genes at different ratios.

As a consequence, **heteroresistance can be a source of false negatives** if the limit of detection of WGS is lower than that of the phenotypic test, as it has been shown in *M. tuberculosis*[46] and *Salmonella enterica*.[47] The limit of detection of WGS is determined by the depth of sequencing coverage. A good example of the need for good quality sequence data (of high enough sequencing depth) is the finding that genotypic resistance sensitivity was 11% and 9% lower for isoniazid and rifampicin respectively, on isolates sequenced at low depth (<10× across 95% of the genome).

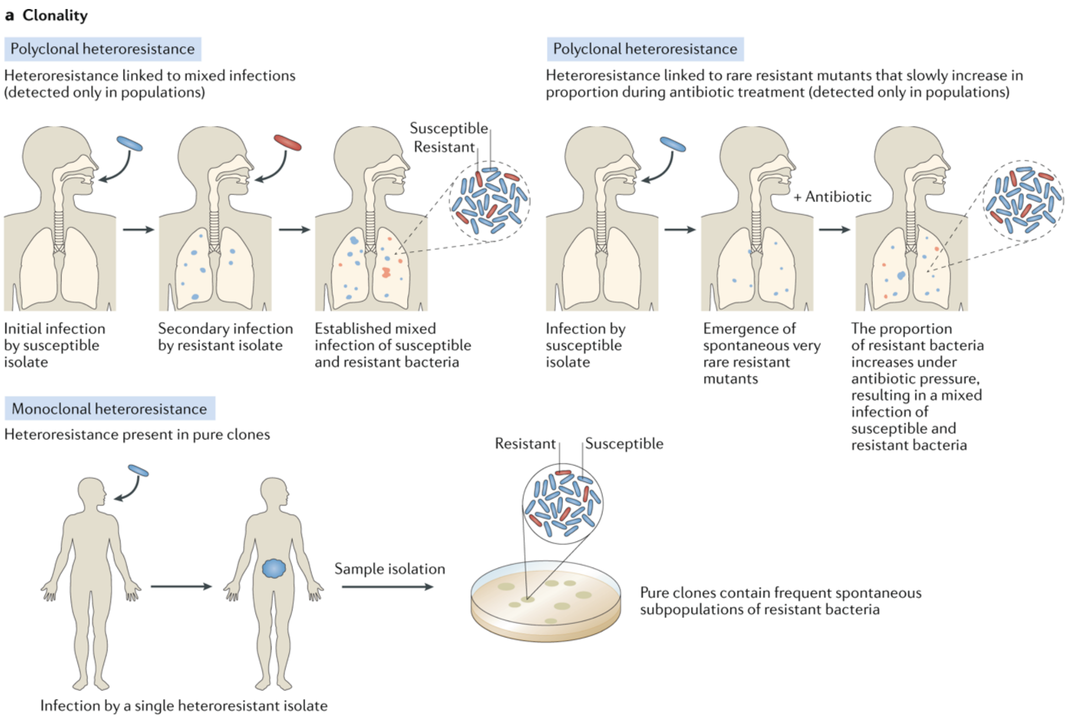

a Clonality
Polyclonal heteroresistance
Heteroresistance linked to mixed infections (detected only in populations)
Susceptible
Resistant
Initial infection by susceptible isolate
Secondary infection by resistant isolate
Established mixed infection of susceptible and resistant bacteria
Polyclonal heteroresistance
Heteroresistance linked to rare resistant mutants that slowly increase in proportion during antibiotic treatment (detected only in populations)
+ Antibiotic
Infection by susceptible isolate
Emergence of spontaneous very rare resistant mutants
The proportion of resistant bacteria increases under antibiotic pressure, resulting in a mixed infection of susceptible and resistant bacteria
Monoclonal heteroresistance
Heteroresistance present in pure clones
Resistant
Susceptible
Sample isolation
Pure clones contain frequent spontaneous subpopulations of resistant bacteria
Infection by a single heteroresistant isolate

*Figure 16* Clonality of heteroresistance

Another source of phenotype-genotype discrepancies is caused by the way antibiotic resistance is categorised into two categories (resistance vs. susceptible) from MIC distributions by applying an epidemiological cut-off or clinical breakpoint (Figure 17).

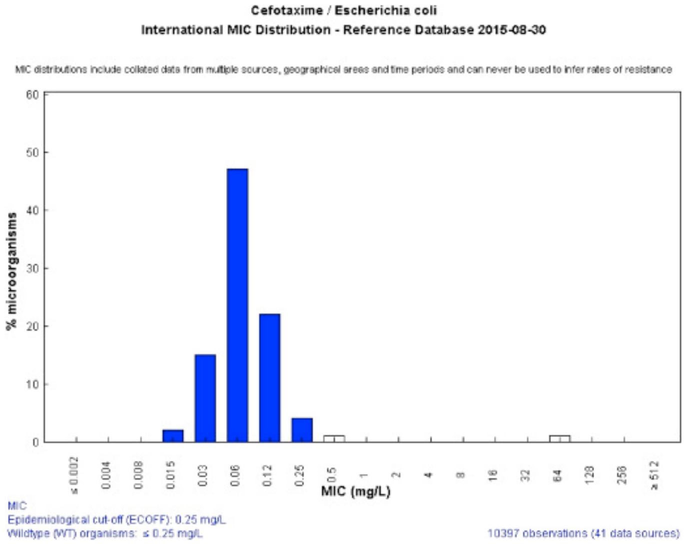


*Figure 17  Cefotaxime MIC distribution for Escherichia coli (n = 10,397 from 41 aggregated distributions).*

For some drugs and organisms, it has been showed that the critical concentrations to distinguish between resistant (expected to carry genetic markers of resistance) and susceptible (expected to be wildtype) strains were originally set too high, leading to the detection of “phenotypically susceptible” strains with ABR markers (i.e. false positives, see example in Figure 18).[48]

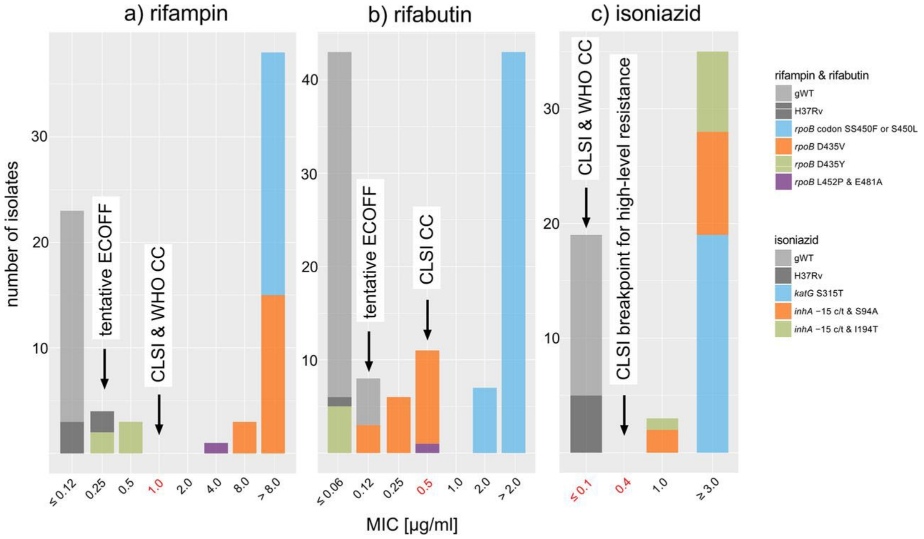


*Figure 18* MIC distributions for rifampin, rifabutin, and isoniazid in *M. tuberculosis*

Another **source of false positives** is the presence of **silenced ABR genes**. Literature reports describe isolates of bacteria that carry an ABR determinant but remain susceptible to the corresponding antibiotic as a consequence of a genetic defect. Despite initial phenotypic susceptibility, such strains represent a source from which antibiotic resistance may re-emerge to cause treatment failure in patients. The prevalence and nature of this phenomenon has been studied in *Staphylococcus aureus*.

And finally, **errors in phenotypic AST** should not be ruled out as a source of phenotype-genotype discrepancies. For example, in a study evaluating ABR WGS-based predictions for *Salmonella* compared with the results of traditional phenotyping assay, the authors found that where initial phenotypic results indicated isolates were sensitive, yet ARGs were detected, repeat phenotypic AST corrected discrepancies.[49] In another study evaluating an automated bioinformatics analysis tool to predict the phenotypic resistance of MRSA, the investigators found that following retesting of discrepant phenotype-genotype results, concordance between phenotypic results and genotypic predictions was 99.69% (Figure 18).

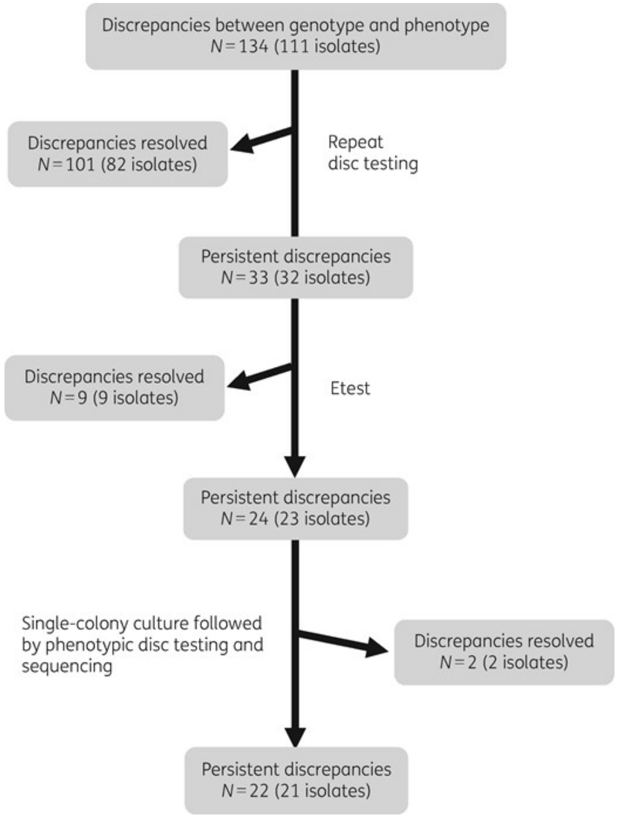
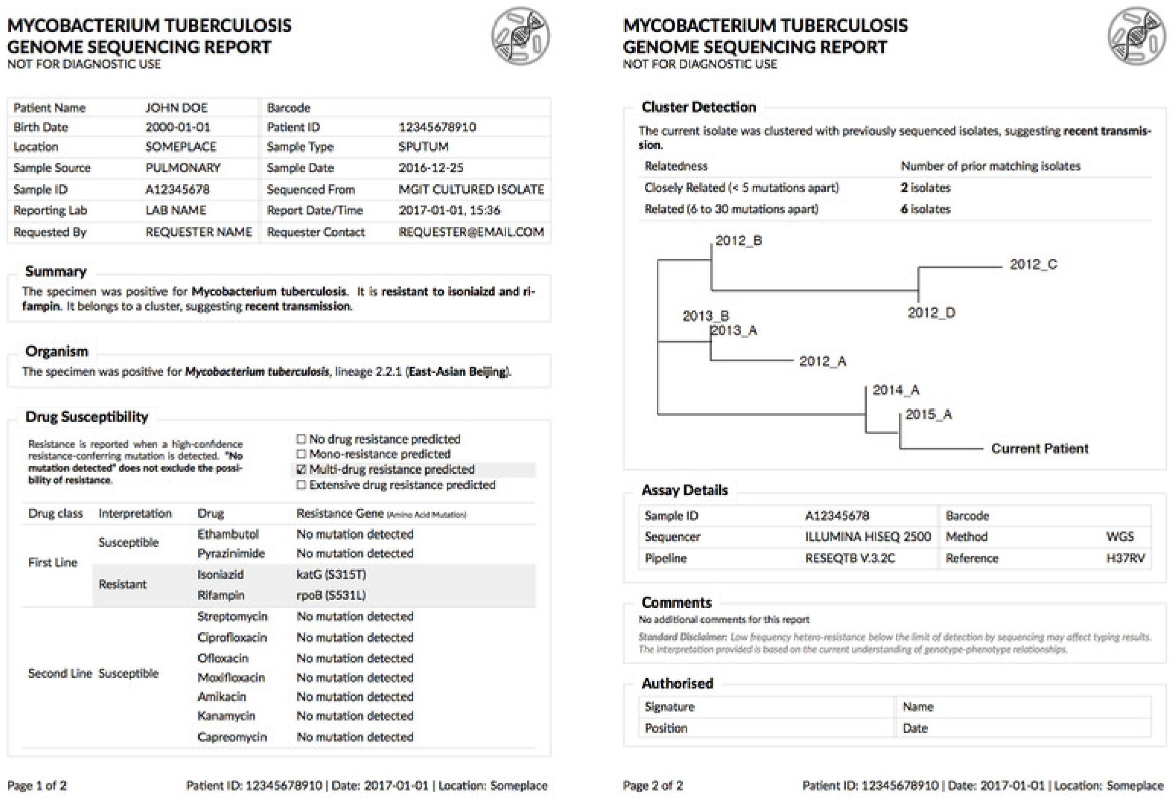


*Figure 19* Algorithm used for retesting of phenotype-genotype discrepancies for MRSA ABR

The detection of **wrongly annotated ABR genetic markers** (i.e. an error in curation) and the presence of an **unknown/novel mechanism of resistance**, should also be considered as potential sources of false positives and false negatives, respectively.

## Genomic reporting standards

Understanding how to report complex genomic test results to stakeholders who may have varying familiarity with genomics - including clinicians, lab technicians, epidemiologists, and researchers - is critical to the successful and sustainable implementation of microbial genome sequencing in clinical and public health laboratories. A few research groups have worked on evidence-based guidelines for designing pathogen genomics reports. A good example of this is an evidence-based design and evaluation of a *M. tuberculosis* whole-genome sequencing clinical report for a reference microbiology laboratory (Figure 20).[50] Another good is the sequence reporting tool developed to detect the infection source for hospital onset COVID-19 infections (HOCIs). The SRT system for prospective use is designed to provide useful and appropriate feedback in both low-incidence and high-incidence settings for new HOCI cases. This is planned through the generation of a concise one-page PDF summary report for each focus sequence (Figure 21). This summary report contains key focus sequence metadata, information regarding the estimated probabilities for infection source and details of up to 10 close sequence matches identified within the same unit/ward and/or elsewhere in the hospital.[51]



**MYCOBACTERIUM TUBERCULOSIS**
**GENOME SEQUENCING REPORT**
NOT FOR DIAGNOSTIC USE

| | | | |
|---|---|---|---|
| Patient Name | JOHN DOE | Barcode | |
| Birth Date | 2000-01-01 | Patient ID | 12345678910 |
| Location | SOMEPLACE | Sample Type | SPUTUM |
| Sample Source | PULMONARY | Sample Date | 2016-12-25 |
| Sample ID | A12345678 | Sequenced From | MGIT CULTURED ISOLATE |
| Reporting Lab | LAB NAME | Report Date/Time | 2017-01-01, 15:36 |
| Requested By | REQUESTER NAME | Requester Contact | REQUESTER@EMAIL.COM |

**Summary**

The specimen was positive for **Mycobacterium tuberculosis**. It is **resistant to isoniaizd and rifampin**. It belongs to a cluster, suggesting **recent transmission**.

**Organism**

The specimen was positive for ***Mycobacterium tuberculosis***, lineage 2.2.1 (**East-Asian Beijing**).

**Drug Susceptibility**

Resistance is reported when a high-confidence resistance-conferring mutation is detected. **"No mutation detected" does not exclude the possibility of resistance.**

☐ No drug resistance predicted
☐ Mono-resistance predicted
☑ Multi-drug resistance predicted
☐ Extensive drug resistance predicted

| Drug class | Interpretation | Drug | Resistance Gene (Amino Acid Mutation) |
|---|---|---|---|
| First Line | Susceptible | Ethambutol | No mutation detected |
| | | Pyrazinimide | No mutation detected |
| | Resistant | Isoniazid | katG (S315T) |
| | | Rifampin | rpoB (S531L) |
| Second Line | Susceptible | Streptomycin | No mutation detected |
| | | Ciprofloxacin | No mutation detected |
| | | Ofloxacin | No mutation detected |
| | | Moxifloxacin | No mutation detected |
| | | Amikacin | No mutation detected |
| | | Kanamycin | No mutation detected |
| | | Capreomycin | No mutation detected |

Page 1 of 2 Patient ID: 12345678910 | Date: 2017-01-01 | Location: Someplace

**MYCOBACTERIUM TUBERCULOSIS**
**GENOME SEQUENCING REPORT**
NOT FOR DIAGNOSTIC USE

**Cluster Detection**

The current isolate was clustered with previously sequenced isolates, suggesting **recent transmission**.

| Relatedness | Number of prior matching isolates |
|---|---|
| Closely Related (< 5 mutations apart) | **2** isolates |
| Related (6 to 30 mutations apart) | **6** isolates |

**Assay Details**

| | | | |
|---|---|---|---|
| Sample ID | A12345678 | Barcode | |
| Sequencer | ILLUMINA HISEQ 2500 | Method | WGS |
| Pipeline | RESEQTB V.3.2C | Reference | H37RV |

**Comments**

No additional comments for this report

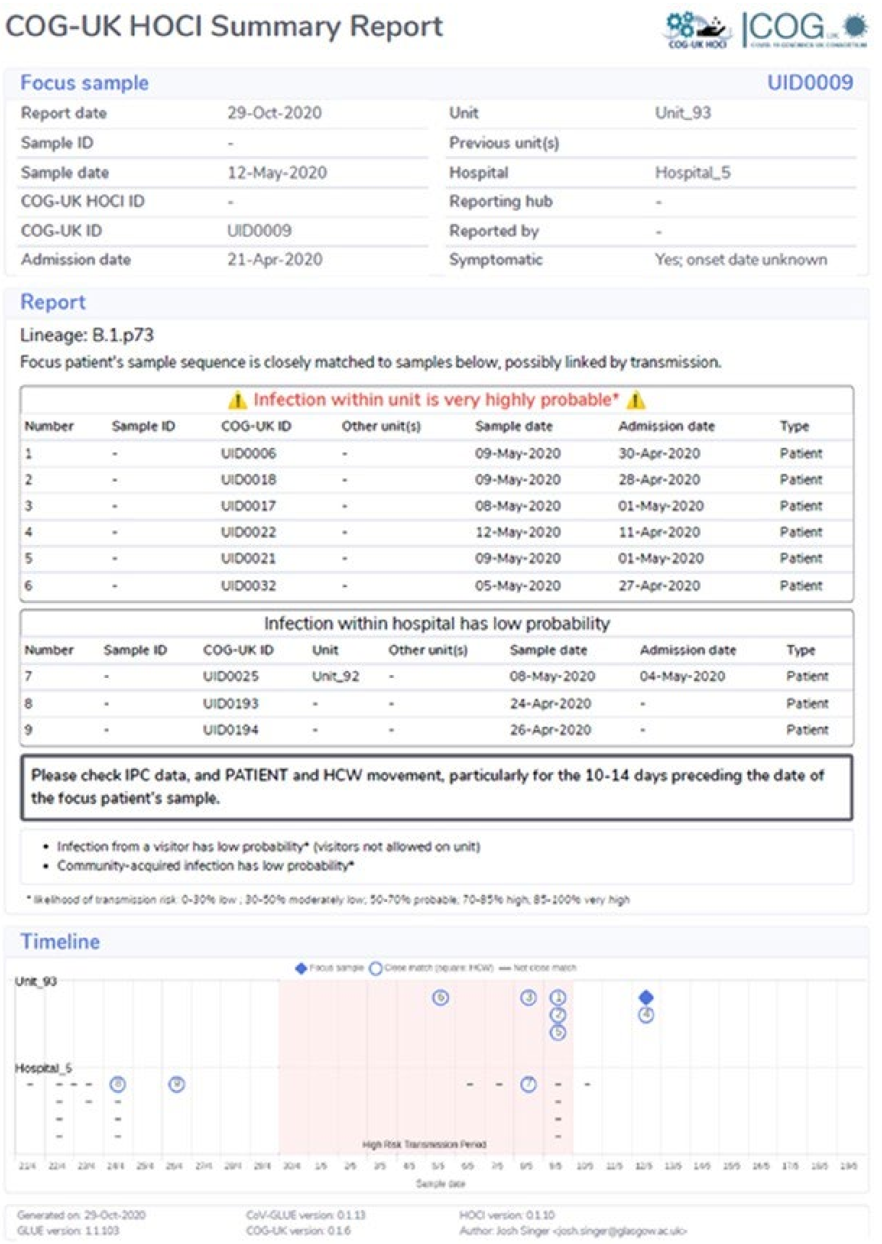
*Standard Disclaimer: Low frequency hetero-resistance below the limit of detection by sequencing may affect typing results. The interpretation provided is based on the current understanding of genotype-phenotype relationships.*

**Authorised**

| | |
|---|---|
| Signature | Name |
| Position | Date |

Page 2 of 2 Patient ID: 12345678910 | Date: 2017-01-01 | Location: Someplace

*Figure 20 Improved design of a M. tuberculosis genomics report*



COG-UK HOCI Summary Report

Focus sample UID0009

| | | | |
|---|---|---|---|
| Report date | 29-Oct-2020 | Unit | Unit_93 |
| Sample ID | - | Previous unit(s) | |
| Sample date | 12-May-2020 | Hospital | Hospital_5 |
| COG-UK HOCI ID | - | Reporting hub | - |
| COG-UK ID | UID0009 | Reported by | - |
| Admission date | 21-Apr-2020 | Symptomatic | Yes; onset date unknown |

Report

Lineage: B.1.p73

Focus patient's sample sequence is closely matched to samples below, possibly linked by transmission.

⚠ Infection within unit is very highly probable* ⚠

| Number | Sample ID | COG-UK ID | Other unit(s) | Sample date | Admission date | Type |
|---|---|---|---|---|---|---|
| 1 | - | UID0006 | - | 09-May-2020 | 30-Apr-2020 | Patient |
| 2 | - | UID0018 | - | 09-May-2020 | 28-Apr-2020 | Patient |
| 3 | - | UID0017 | - | 08-May-2020 | 01-May-2020 | Patient |
| 4 | - | UID0022 | - | 12-May-2020 | 11-Apr-2020 | Patient |
| 5 | - | UID0021 | - | 09-May-2020 | 01-May-2020 | Patient |
| 6 | - | UID0032 | - | 05-May-2020 | 27-Apr-2020 | Patient |

Infection within hospital has low probability

| Number | Sample ID | COG-UK ID | Unit | Other unit(s) | Sample date | Admission date | Type |
|---|---|---|---|---|---|---|---|
| 7 | - | UID0025 | Unit_92 | - | 08-May-2020 | 04-May-2020 | Patient |
| 8 | - | UID0193 | - | - | 24-Apr-2020 | - | Patient |
| 9 | - | UID0194 | - | - | 26-Apr-2020 | - | Patient |

**Please check IPC data, and PATIENT and HCW movement, particularly for the 10-14 days preceding the date of the focus patient's sample.**

- Infection from a visitor has low probability* (visitors not allowed on unit)
- Community-acquired infection has low probability*

* likelihood of transmission risk: 0-30% low ; 30-50% moderately low; 50-70% probable; 70-85% high; 85-100% very high

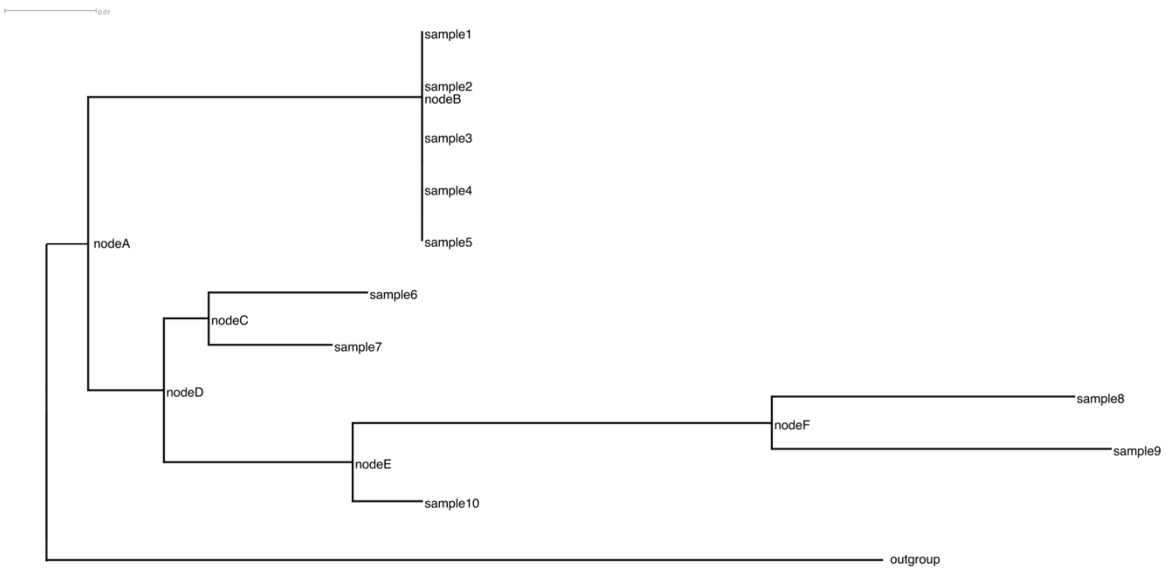
Timeline

Generated on: 29-Oct-2020 CoV-GLUE version: 0.1.13 HOCI version: 0.1.10

GLUE version: 1.1.103 COG-UK version: 0.1.6 Author: Josh Singer <josh.singer@glasgow.ac.uk>

*Figure 21 Sequence reporting tool (SRT) for hospital onset COVID-19 infections (HOCIs)*

## Appendix 1: Exercises on interpreting phylogenetic trees



Question 1. based on the tree above, what internal node corresponds to the most recent common ancestor of samples 8 and 10:

- Node F
- Node D
- Sample 7
- Node E

Question 2. Based on the tree above, which group of samples are most closely related:

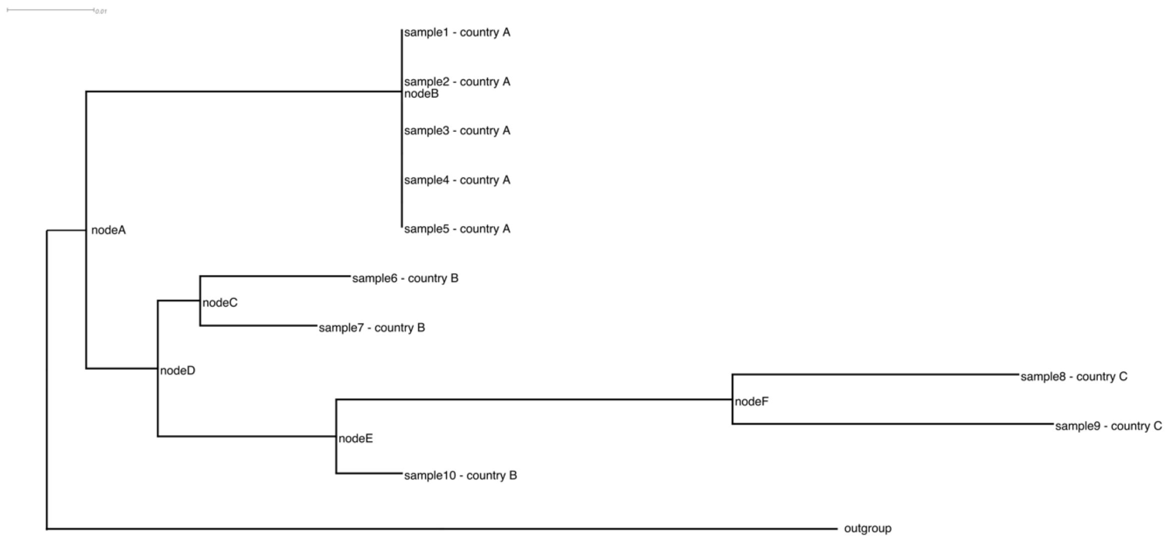
- Samples 1 to 5
- Samples 6 & 7
- Samples 6 to 10
- Samples 8 & 9

Question 3. Based on the tree above, which of the following statements referring to sample 10 is more accurate:

- Sample 10 is more closely related to sample 7 than to sample 8
- Sample 10 is more closely related to sample 8 than to sample 7
- Sample 10 is equally related to sample 7 and sample 8
- Sample 10 is related to sample 8, but it is not related to sample 7

Question 4. Based on the tree above, which of the following statements referring to sample 7 is more accurate:

- Sample 7 is more closely related to sample 8 than to sample 10
- Sample 7 is more closely related to sample 10 than to sample 8
- Sample 7 is equally related to sample 8 and sample 10
- Sample 7 is related to sample 8, but it is not related to sample 10



Question 5. Based on the country of origin of samples on the tree above, which of the following statements about transmission events is more certain:

- The common ancestor of samples 6 to 10 (node D) most likely circulated in country A first and later on transmitted to country B and C
- The common ancestor of samples 6 to 10 (node D) most likely circulated in country B first and later on transmitted to country C
- The common ancestor of samples 6 to 10 (node D) most likely circulated in country C first and later on transmitted to country B
- The common ancestor of samples 6 to 10 (node D) could have circulated in country A or B

Question 6. Based on the country of origin of samples on the tree above, which of the following statements about transmission events is more certain:

- The common ancestor of samples 1 to 10 (node A) most likely circulated in country A first and later on transmitted to country B and C
- The common ancestor of samples 1 to 10 (node A) most likely circulated in country B first and later on transmitted to country A and C
- The common ancestor of samples 1 to 10 (node A) most likely circulated in country C first and later on transmitted to country A and B
- The common ancestor of samples 1 to 10 (node A) could have circulated in country A or B