T3connect

How to design and deliver pathogen genomics training for health and research professionals

Module 3D
How to train - data interpretation and applications
08/03/23

wellcome connecting science

Centre for Genomic Pathogen Surveillance

BIG DATA INSTITUTE

UNIVERSITY OF OXFORD

Content

Table 1 Topics and sub-topics of teaching content

Teaching topic	Sub-topics
Genomic QC metrics	QC metrics at different sequence analysis stages
	Thresholds for quality metrics
	Controls and validated QC procedures
	Detecting contamination
Speciation and strain typing	Ribosomal MLST
	Taxonomic classifiers
	Strain typing at different resolutions: MLST, core-genome MLST and whole-genome MLST
	Lineage-specific markers
Phylogenetic trees interpretation	Basics of phylogenetic tree reconstruction
	Extracting strain relatedness information from trees
	Area of applications: foodborne, hospital, community outbreaks and STI outbreaks (e.g. TB)
Visualisation of genomic and epidemiological data	Annotated trees
	Specialised tools: MicroReact, Nextstrain
	Patient timeline plots
Genetic relatedness thresholds	How thresholds are applied and interpreted
WGS-based AMR prediction	Early proof-of-concept studies
	Available approaches, databases and tools
	Diagnostic accuracy of genotypic determinations
	Sources of genotype-phenotype discrepancies
Genomic reporting standards	Pathogen genomics reports

Strategies to deliver topics and sub-topics of pathogen genomics content

Done: View

Examples of strategies

Done: View

Group activity: Design a session on data interpretation and applications

Done: View

wellcome connecting science

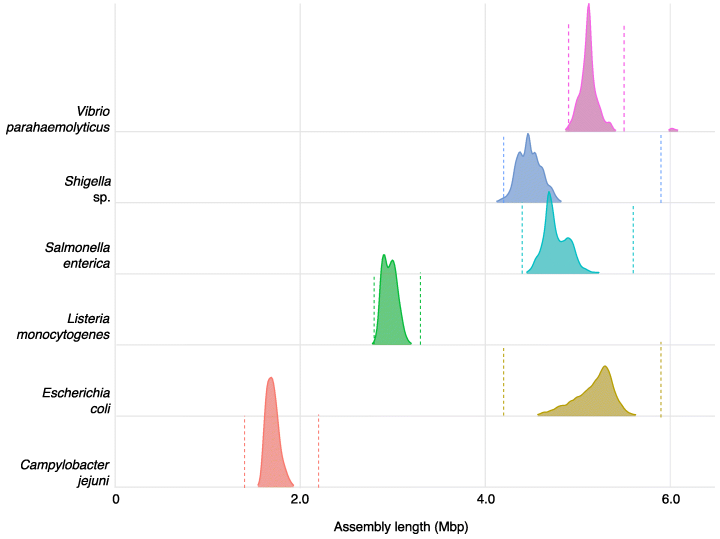
T3connect

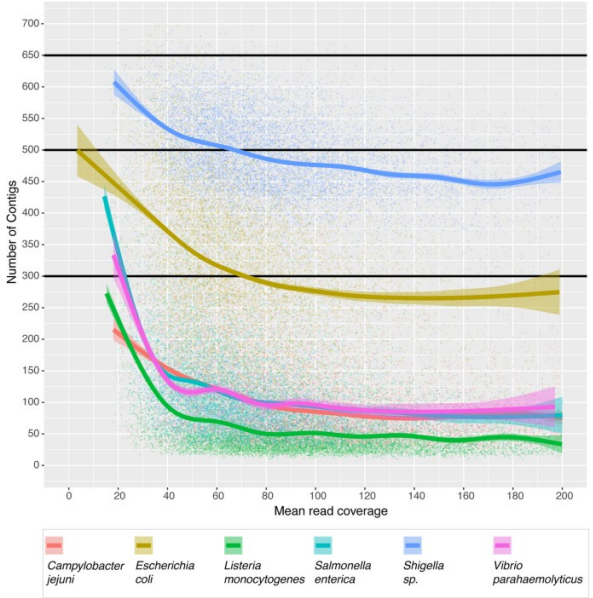
Centre for Genomic Pathogen Surveillance

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

Thresholds for quality metrics should be set beforehand, which are often organism specific.





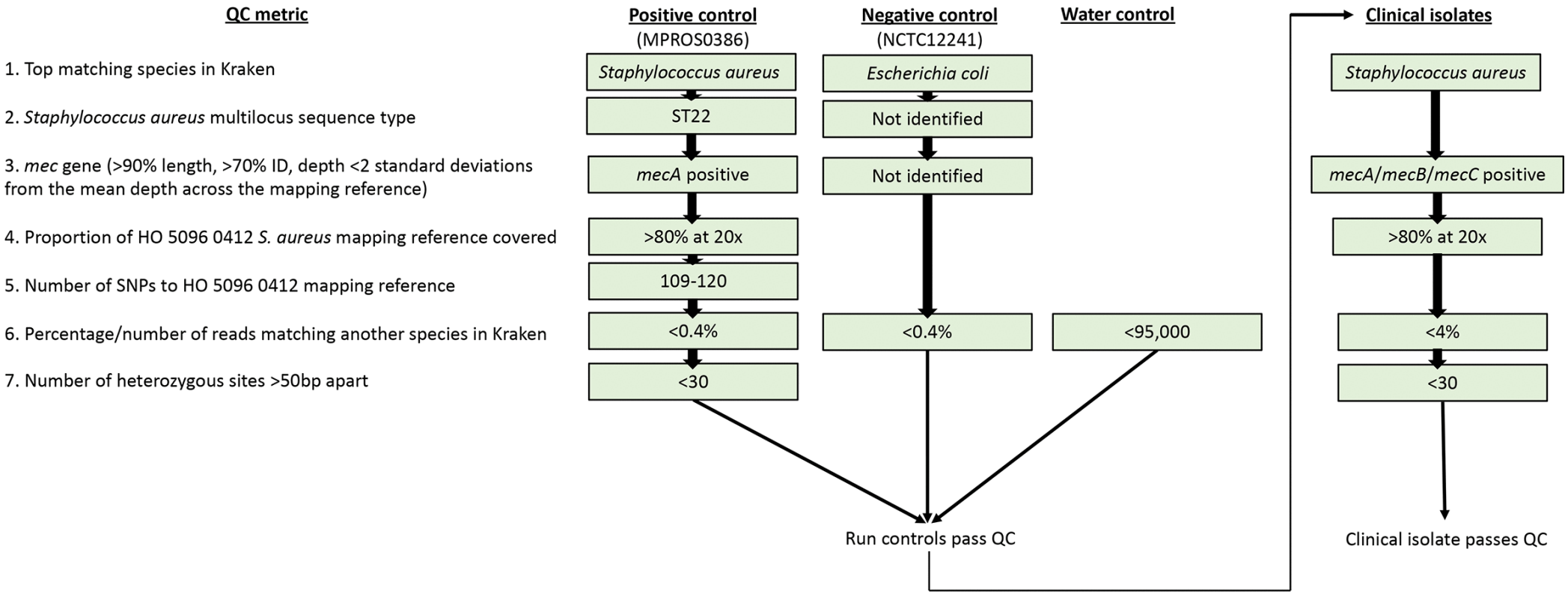
wellcome connecting science

T3connect

Centre for Genomic Pathogen Surveillance

Interpretation of genomic QC metrics

Example: QC flowchart for passing/failing controls and clinical isolates for MRSA sequencing



wellcome connecting science

T3connect

Centre for Genomic Pathogen Surveillance

Interpretation of genomic QC metrics: teaching strategies

Table 2. Teaching strategies and assessment

Topic	Teaching strategies	Assessment
Genomic QC metrics	Collect examples of problematic samples or sequencing batches at your institution; what genomic metrics were used to identify bad quality genomes?	Provide learners with a mixture of the real-world good and bad quality samples/genomes. This may include raw sequencing data, processed sequence data and/or final genomic reports.
	What information (i.e. combination of various genomic QC metrics) helped diagnose what went wrong in the upstream data collection, processing and/or sequencing steps?	Based on the metrics that did not pass pre-defined QC thresholds, ask learners to identify the error and stage in sample processing (e.g. specimen culture, DNA extraction, sequencing run) that may have led to a bad quality sample or batch.
	Impact of bad quality samples on interpretation	Provide learners with case studies on wrong interpretation, and wrong clinical/epidemiological actions that would have followed, caused by bad-quality samples; and how interpretation changed once bad sample(s) were removed.
	Stress key concepts in genomic QC. For example: different sources of contamination (different species vs. strain contamination); how QC thresholds are set; the type of controls used; QC thresholds may vary by microbial organism.	Assess these concepts by selecting a diverse set of bad-quality samples

wellcome connecting science

T3connect

Centre for Genomic Pathogen Surveillance

Introduction to phylogenetic trees

How are **phylogenetic trees reconstructed** from the number and pattern of shared mutations between strains (and assumptions)

Introduce **phylogenetic nomenclature**, as terms like “clade”, “tips”, “topology” or “branches” are commonly used in the field of ID genomic.

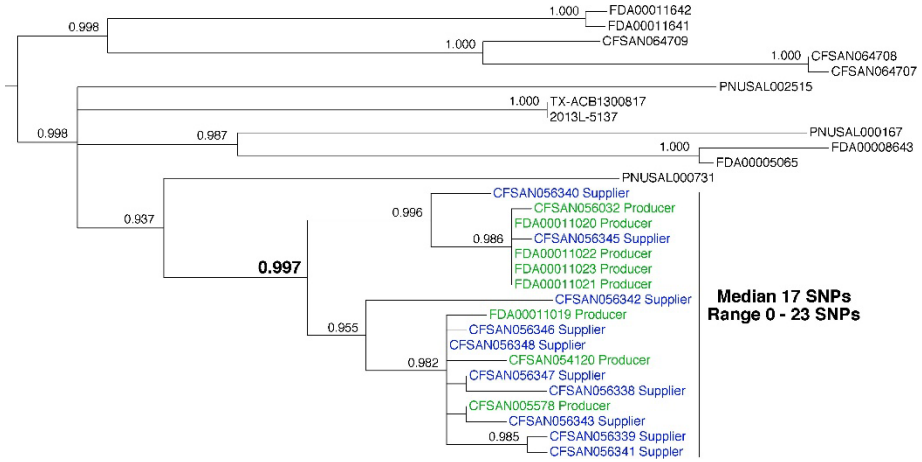
Online resources on how to read phylogenetic trees that introduce these phylogenetic concepts and nomenclature including.

- The EBI course on phylogenetics, for example, places an emphasis on how to read and interpret phylogenetic trees
- 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.

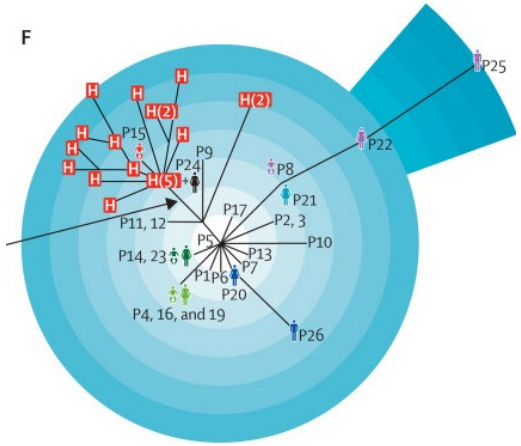
Interpretation of phylogenetic trees for ID epidemiology

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 case studies



Phylogenetic analysis of *Listeria monocytogenes* isolated from ice cream samples



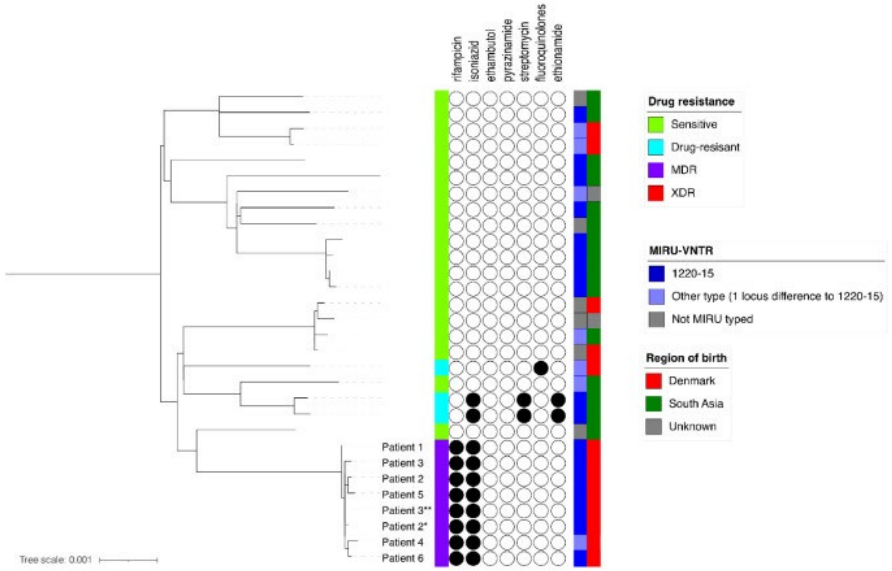
Phylogeny of the MRSA SCBU outbreak

wellcome connecting science

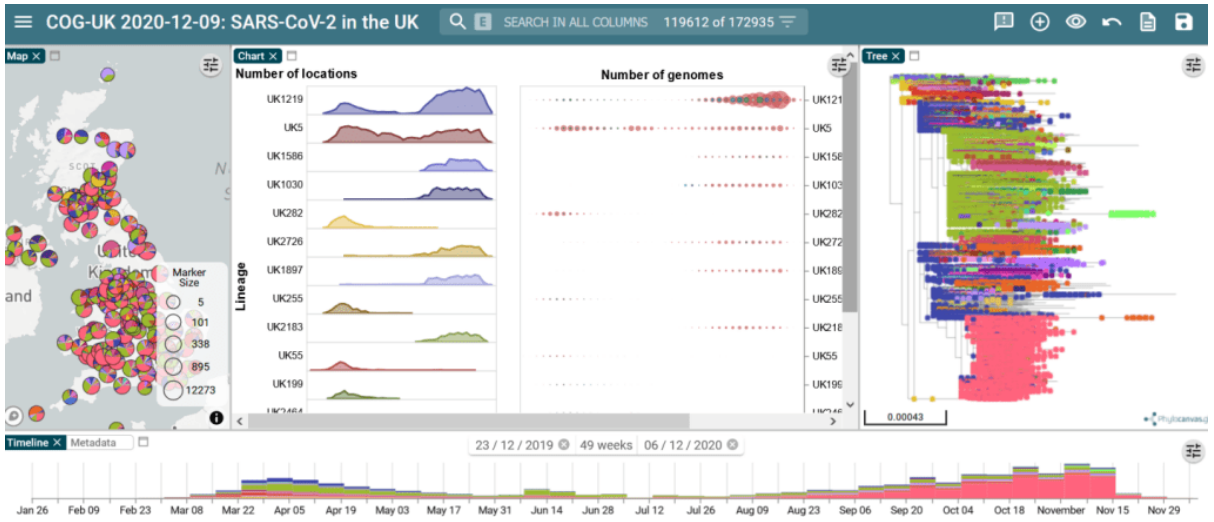
T3connect

Centre for Genomic Pathogen Surveillance

Visualisation of genomic and epidemiological data



iTOL



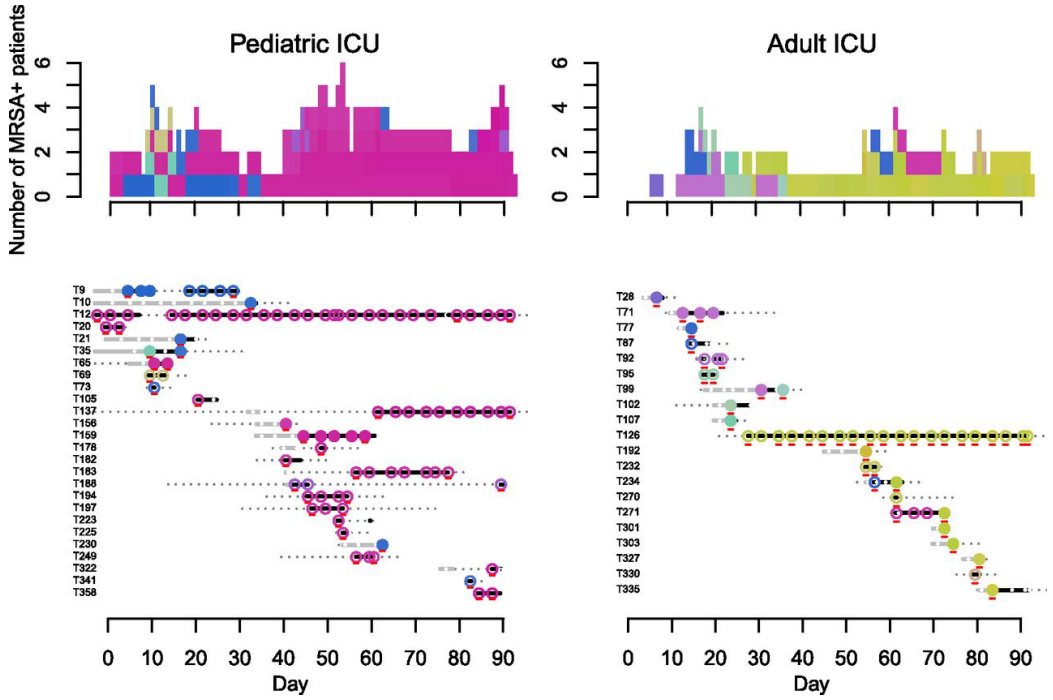
MicroReact

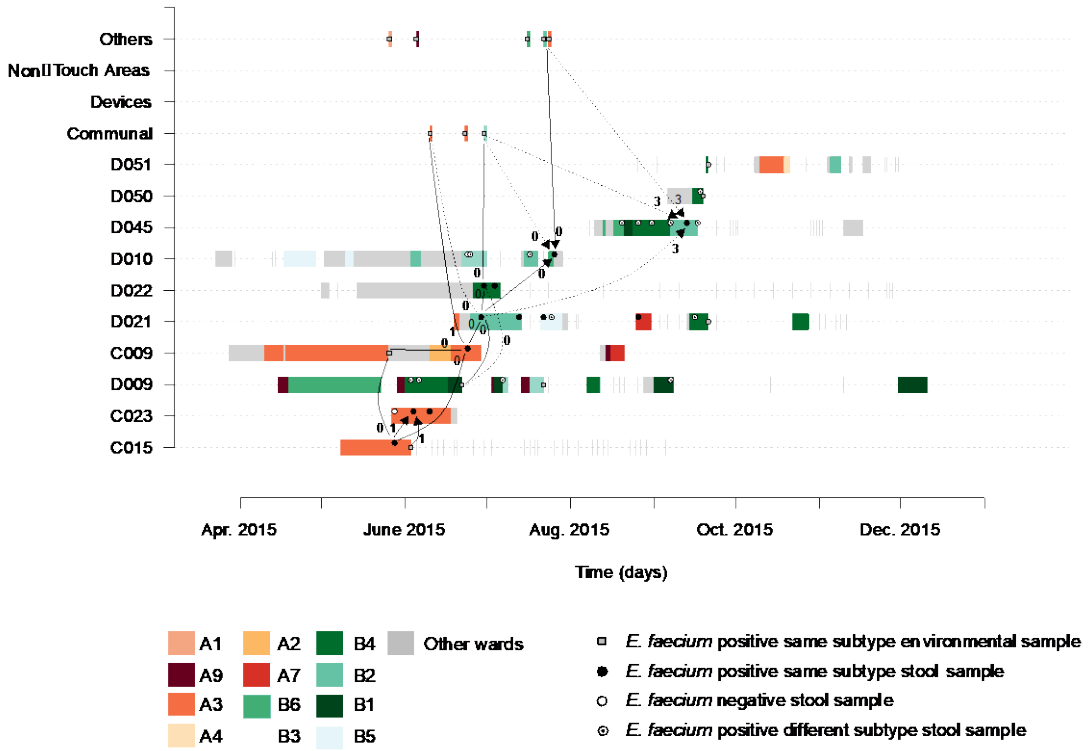
wellcome connecting science

T3connect

Centre for Genomic Pathogen Surveillance

Visualisation of genomic and epidemiological data





wellcome connecting science

T3connect

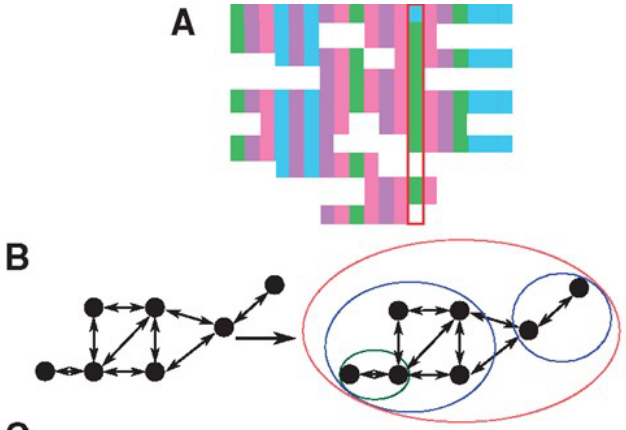
Centre for Genomic Pathogen Surveillance

Genetic relatedness thresholds

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

It is increasingly acknowledged that epidemiological follow-up (i.e. detection of common epidemiological links) is needed to confirm definite transmission.

Limitations of the SNP cut-off approach



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

SNP address

wellcome connecting science

T3connect

Centre for Genomic Pathogen Surveillance

Genetic relatedness thresholds: teaching strategies

Genetic relatedness thresholds	Introduce how genetic relatedness thresholds (SNP cut-offs) are applied and interpreted to identify pathogen transmission from genomic data.	As explained above, use a variety of genomic epidemiology case-studies that applied genetic relatedness thresholds to detect transmission clusters, rule out transmission and guide epidemiological investigations.
	Introduce concepts commonly used in genomic epidemiology: e.g. transmission cluster, genetic link, weak vs. strong epidemiological link, hospital vs. community epidemiological link, etc.	Reenforce concepts commonly used in genomic epidemiology.
	Introduce approaches used to determine SNP cut-offs: based on the maximum within-host diversity or the distribution of genetic distances between strains from cases with confirmed epidemiological links.	Put an emphasis on limitations and strengths of SNP cut-offs, and give example on how the identification of common epidemiological links are still essential to confirm definite transmission in genomic epidemiology investigations.

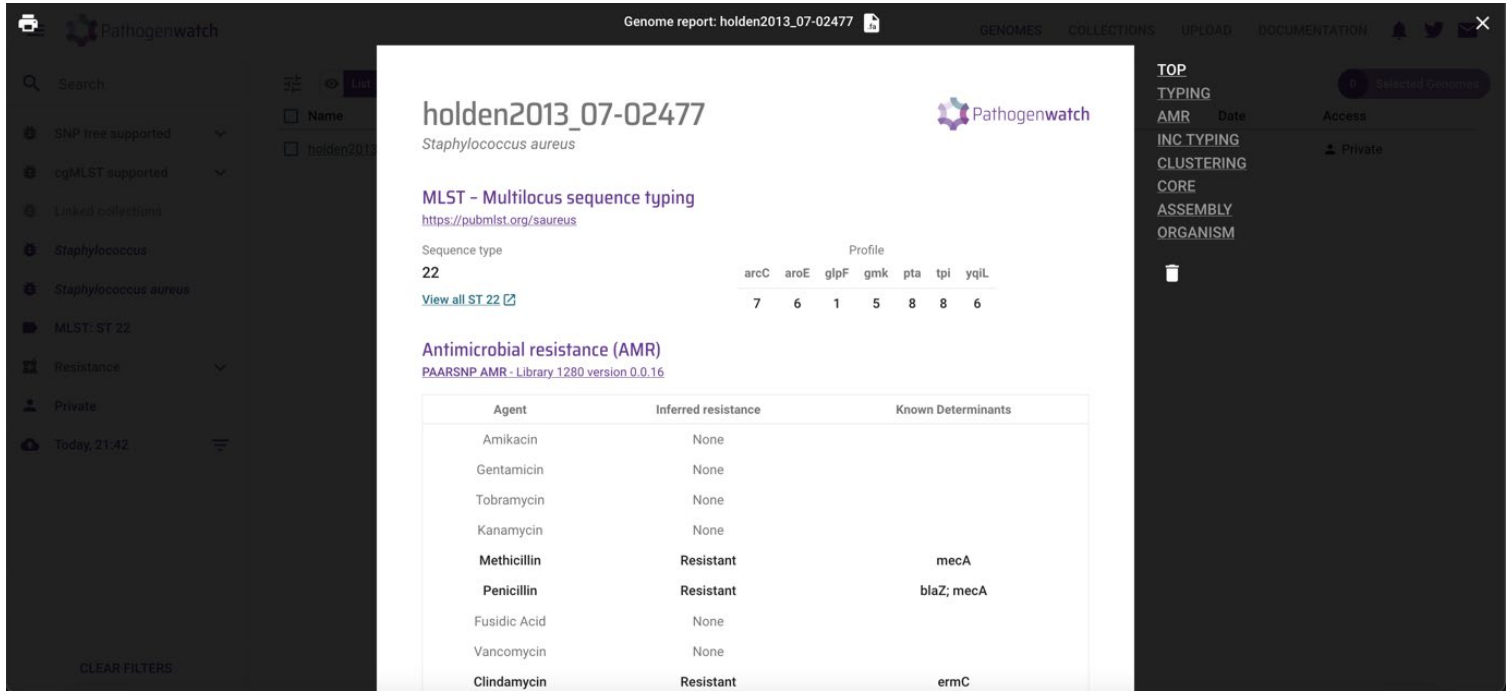
Interpreting genotypic AMR predictions: teaching strategies

WGS-based AMR prediction	Introduce key biological, evolutionary and genetic concepts driving the action of antimicrobial drugs and causes of antimicrobial resistance in microbial organisms. For example: acquisition of new AMR genes via horizontal-gene transfer (HGT), acquisition of genetic variants in existing regions of the core or accessory genome due to mutation and recombination, etc.	There are plenty of online courses, resources and scientific reviews covering mechanisms of action of antibiotics and mechanisms of AMR. A few examples include: - Darby, E. M. et al. Molecular mechanisms of antibiotic resistance revisited. Nature Reviews Microbiology 1–26 (2022) doi:10.1038/s41579-022-00820-y. - Boolchandani, M., *et al.* Sequencing-based methods and resources to study antimicrobial resistance. *Nature Reviews Genetics* 20, 356–370 (2019). - The Whys and Wherefores of Antibiotic Resistance: http://perspectivesinmedicine.cshlp.org/content/7/2/a025171.full
	Present early proof-of-concept studies demonstrating that, in principle, whole-genome sequencing can be as sensitive and specific as phenotypic methods at predicting antimicrobial resistance.	The datasets and examples of early proof-of-concept studies in *Staphylococcus aureus,*[1] *Mycobacterium tuberculosis*,[2] *Escherichia coli* or *Klebsiella pneumoniae*[3] can be used to exemplify the use of WGS to predict AMR, and to give a historical context.
	List available approaches, databases and bioinformatic tools to predict AMR from genomic sequences.	Online and command-line 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 AMR, which also host underlaying curated databases of AMR genetic markers needed to make these predictions. Teaching materials using these tools can be designed that make use of real-world genomic data, extracted from scientific papers or from your own institution.

Available approaches, databases and tools

Most common approach is the look-up table or rule-based approach

Tools like AMRFinder, CARD RGI, ResFinder, or Pathogenwatch are among the most commonly used bioinformatic tools to determine ABR from WGS



wellcome connecting science

T3connect

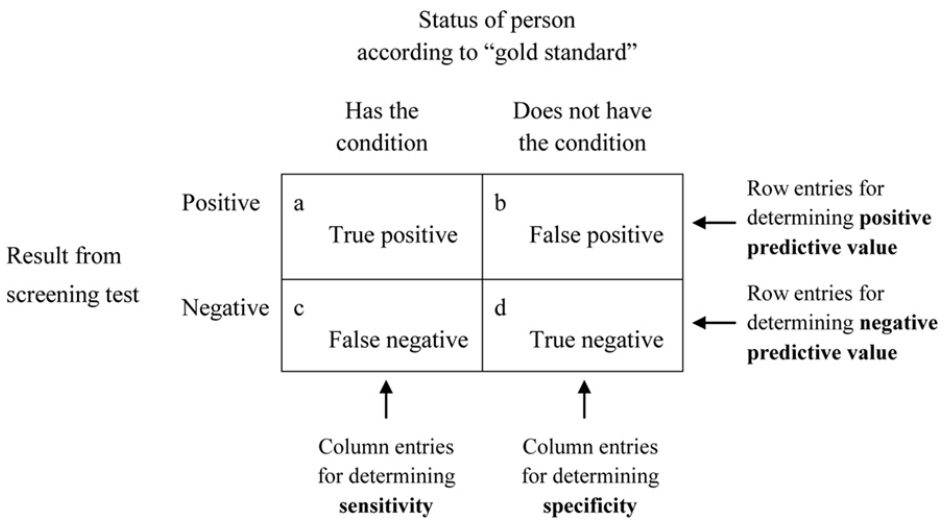
Centre for Genomic Pathogen Surveillance

Interpreting genotypic AMR predictions: teaching strategies

WGS-based AMR prediction	Introduce diagnostic metrics and approaches used to assess the accuracy of genotypic determinations with population-based studies.	It is important to stress that the accuracy of AMR genotypic determinations needs to be assessed with population-based studies; and that this may differ by antimicrobial and microbial species.
	Explain the limitations of WGS-based determination of AMR and sources of genotype-phenotype discrepancies	Provide learners with a mixture of the real-world strains/genomes with matching and incongruent AMR genotype-phenotypes. This may include raw sequencing data, processed sequence data and/or final genomic reports, along with phenotypic AST results for comparisons. Cases may include: bad quality genomes (e.g. with contamination) leading to a wrong genotypic AMR prediction, clonal hetero-resistance, mixed infections, strains with silenced AMR genes, etc.

Diagnostic accuracy of ABR genotypic determinations

The accuracy of genotypic predictions should be assessed for individual antibiotics and bacterial species.



$$\text{Sensitivity} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

$$\text{Specificity} = \frac{\text{True Negatives}}{\text{True Negatives} + \text{False Positives}}$$

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 (96.5–97.7)	99.0 (98.7–99.2)
Rifampin	2743	69	7	84	2903	85	6763	232	147	7227	97.5 (96.9–98.1)	98.8 (98.5–99.0)
Ethambutol	1410	81	94	55	1640	468	6835	781	70	8154	94.6 (93.3–95.7)	93.6 (93.0–94.1)
Pyrazinamide	863	82	117	77	1139	204	6146	197	108	6655	91.3 (89.3–93.0)	96.8 (96.3–97.2)

Walker AS *et al*. NEJM. 2018;379(15).

wellcome connecting science

T3connect

Centre for Genomic Pathogen Surveillance

Genomic reporting standards

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

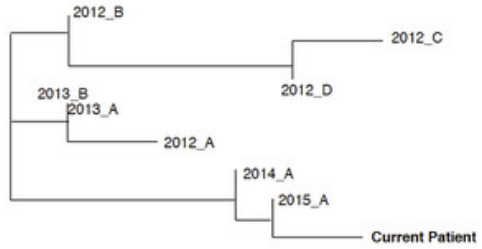
Page 1 of 2 Patient ID: 12345678910 | Date: 2017-01-01 | Location: Someplace

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

Crisan *et al.* PeerJ (2018)

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

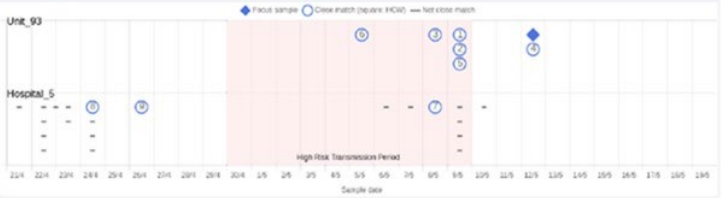
Number	Sample ID	COG-UK ID	Unit	Other unit(s)	Sample date	Admission date	Type
7	-	UID0025	Unit_92	-	08-May-2020	04-May-2020	Patient
8	-	UID0193	-	-	24-Apr-2020	-	Patient
9	-	UID0194	-	-	26-Apr-2020	-	Patient

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>

Stirrup *et al.* eLife (2021)

wellcome connecting science

T3connect

Centre for Genomic Pathogen Surveillance

Resources

Strategies to deliver topics and sub-topics of pathogen genomics content

Done: View

Examples of strategies

Done: View

Group activity: Design a session on data interpretation and applications

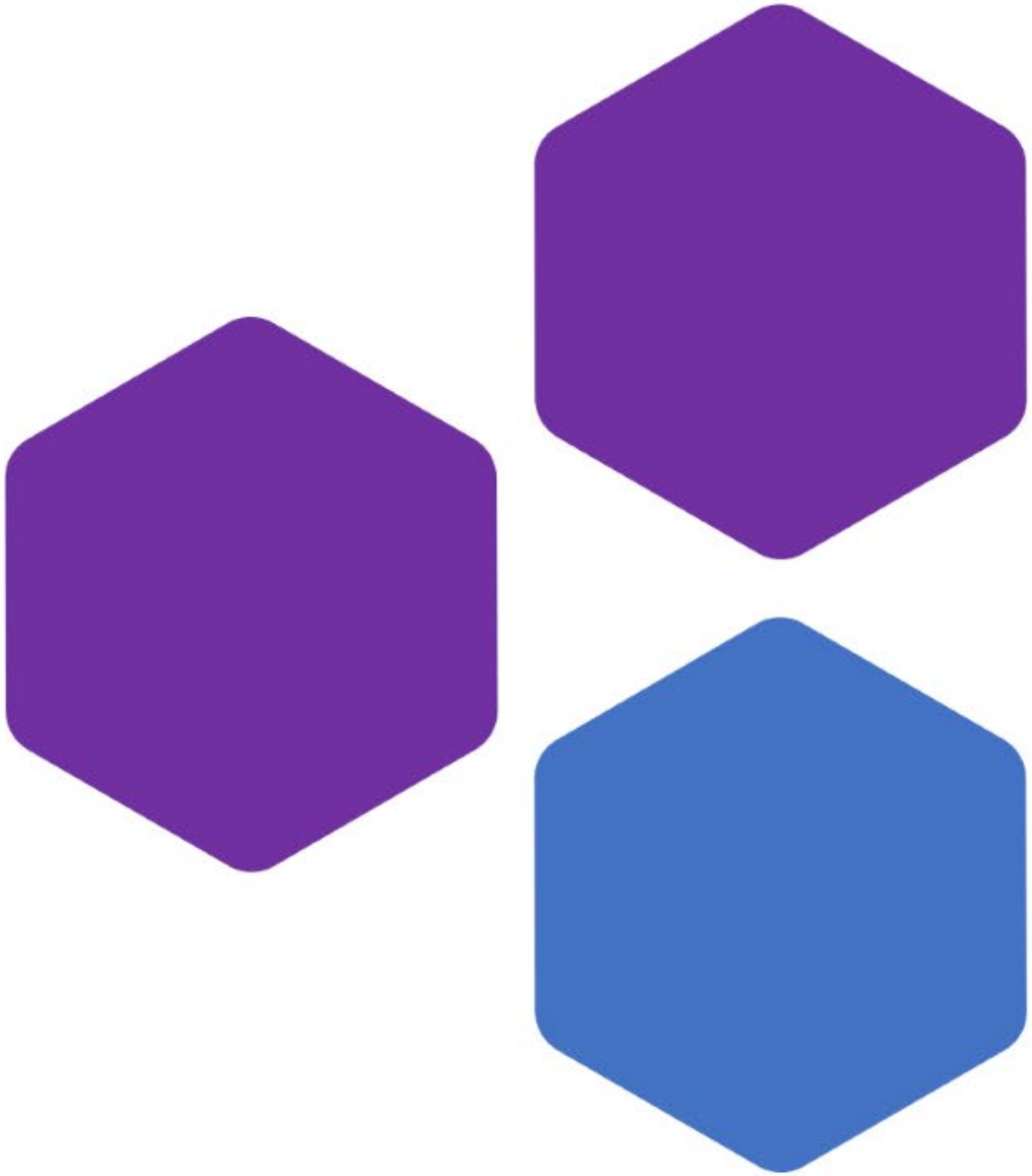
Done: View

wellcome connecting science

T3connect

Centre for Genomic Pathogen Surveillance

Thank you



T3connect