Topics and sub-topics of teaching content

The table below includes a list of topics and sub-topics in the broader area of data interpretation in pathogen genomics. While not aimed to be a comprehensive list, this table gives you an example of topical areas to cover in a hypothetical course, and an example of structure on how to identify sub-topics within each teaching topic.

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 content

The table below gives examples on how each topic on data interpretation can be delivered (strategies) and assessed:

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
Speciation and strain typing	Identify the clinical and public health scenarios/applications where speciation is key (e.g. aetiological diagnosis of infections using metagenomics)	Present metagenomic case-studies in the scientific literature and/or at your own institution where speciation from genomic data was needed to diagnose the aetiological microbe of an infection. Here, standard laboratory practises already in use to

		diagnose infections (e.g. bacterial culture or gram-staining) could be presented.
	Identify the stage(s) in the genome sequence analysis pipeline where speciation is key (e.g. confirmation of target cultured microbe, identification of contamination)	Provide learners with a mixture of real-world good and bad quality samples/genomes, that is, containing the expected target microbial organism or contamination with other species, respectively. This may include exercises using raw sequencing data, processed sequence data and/or final genomic reports. Give examples of actions that would follow after identification of contamination.
	Describe the different available speciation approaches and tools (ribosomal MLST, average nucleotide identity methods, taxonomic classifiers, genetic markers, etc.) Teach the basic biological and genetics concepts underlying the speciation approach (e.g. conservation of ribosomal genes, particular genetic markers, etc.).	See teaching strategies above for 'Speciation and strain typing'. Apply and interpret the results of different available speciation tools to the same samples/genomes. In principle, different tools should give complementary and confirmatory results of the microbial organisms present in the sequenced sample. The bioinformatics approach and tool used to confirm the target organism may differ by microbial organism.
	Present the different approaches, bioinformatics tools, databases and schemes available for genotyping microbial strains: e.g. MLST, core-genome MSLT and whole-genome MLST.	Apply genotyping tools/schemes of different resolutions to the same group of strains/genomes, and identify the appropriate tool/scheme and added value of this based on the genotyping resolution required for different applications.
	Identify the research, clinical and public health scenarios/applications where strain-typing is employed (e.g. outbreak investigation), and what resolution (e.g. core vs. whole-genome MLST) is required for different applications.	Present case studies where strain typing can be employed for genomic QC purposes: e.g. identifying cases of mixed infection or contamination with different strains.
	Present the genotyping approaches/tools/schemes that are organism	Identify scenarios, case studies and datasets where organism-specific tools would be more appropriate: e.g. TB-Profiler for *M.*

	agnostic (e.g. MLST) versus those that are designed for specific microbial organism(s).	*tuberculosis* (https://github.com/jodyphelan/TBProfiler) or GenoTyphi for *Salmonella* Typhi). (https://github.com/katholt/genotyphi)
Phylogenetic trees interpretation	Introduce the basic concepts of phylogenetics, including nomenclature ("clade", "tips", "topology", "branches", etc.) and assumptions of phylogenetic tree reconstructions (e.g. clonal reproduction, mutation rates.) 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, public-health reports and literature reviews that can be used to develop teaching materials on this topic.	Use available online resources on how to read phylogenetic trees that introduce these phylogenetic concepts and nomenclature, especially those placing an emphasis on reading and interpreting phylogenetic trees in the context of infectious diseases epidemiology. Include exercises on how to interpret phylogenetic trees, with a particular focus on extracting strain relatedness information (see Appendix 1 for an example). Reading phylogenetic trees correctly may be relatively straightforward for an expert user, but should not be taken for granted. Use genomic epidemiology case-studies to identify and understand the source of foodborne (e.g. *Salmonella* or *Listeria* outbreaks); hospital (e.g. MRSA or *Pseudomonas*); community (e.g. TB) or sexually transmitted outbreaks.
Visualisation of genomic and epidemiological data	Visualising and annotating phylogenetic trees with epidemiological data (e.g. patient identifiers, collection times, geographical locations, etc.) are a simple and powerful approach to intuitively and visually investigate outbreaks.	Make use of publicly-available tools such as Microreact or iTOL (https://itol.embl.de/) to visualise phylogenetic trees along with annotations of epidemiological data, when writing and delivering teaching materials.
	Introduce more specialised tools such as Microreact (https://microreact.org/) or Nextstrain (https://nextstrain.org/) which	Make use of publicly-available tools such as Microreact or Nextstrain when writing and delivering teaching materials on

	include purpose-built functionality for genomic epidemiology investigations.	interpreting phylogenetic trees and epidemiological data in genomic epidemiology investigations; as explained above.
	Introduce the variety of plots and data visualisation strategies used to visualise genomic and epidemiological data (e.g. patient timelines, networks, minimum spanning tree, blobograms, etc.)	Extract the variety of plots and data visualisation strategies used in real-world genomic epidemiology investigations. For example, timelines of patients visiting different hospital wards. Put an emphasis of how to interpret these plots, e.g. to identify common epidemiological markers (e.g. patients staying at the same hospital ward); transmission patters such as super-spreaders based on the topology of the tree or shape of minimum spanning trees; etc.
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.
WGS-based AMR prediction	Introduce key biological, evolutionary and genetic concepts driving the action of antimicrobial drugs and causes of antimicrobial resistance in microbial	There are plenty of online courses, resources and scientific reviews covering mechanisms of action of antibiotics and mechanisms of AMR. A few examples include:

	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.	- 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.
	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.
Genomic reporting standards	Introduce the importance of designing and evaluating evidence-based guidelines and standards for pathogen sequencing clinical and public health reports.	Introduce examples of evidence-based pathogen sequencing reports, for clinical and public health applications, extracted from the scientific literature and implemented in healthcare and public health institutions.
	Identify the key patient metadata and genomic data fields required to make clinical and public health actions	Present cases of standardisation and accreditation (e.g. ISO certification) of genomic pipelines.[7]