Day 3: Design Training

George Githinji, Kirsty Lee Garson, Aquillah Kanzi, Stanford Kwenda, Alice Matimba, Leigh Jackson, Amadou Diallo

Please add your institution logo to the slide theme

Session outline

This Session will cover tools and approaches to designing activities for effective training in data analysis and interpretation using a case study on data interpretation and applications

- Concepts of data interpretation
- Strategies used to teach these concepts

Session outcomes

At the end of this module participants will be able to:

- Identify knowledge, concepts and skills required for interpretation pathogen genomics data
- Determine appropriate topics and sub-topics for training in data interpretation.
- Outline training activities which apply active learning strategies.
- Identify appropriate tools and resources required to deliver training in data interpretation

Developing modules - Pathogen Genomics Data Interpretation - Step 1

- Genomic QC metrics
- Speciation and strain typing
- Phylogenetic trees interpretation
- Visualisation of genomic and epidemiological data
- Genetic relatedness thresholds
- WGS-based AMR prediction
- Genomic reporting standards

Developing modules - Step 2

Teaching topic (Module)	**Sub-topics in module**
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

Design training strategies and activities - Step 3

1. Refine sub-topics, objectives and learning outcomes
2. Training strategies and possible activities
3. Assessment
4. (Resources and tools)
5. (Feedback strategy)

Developing Training in Bioinformatics

Training bioinformatics users

Making the variety and complexity of bioinformatics more accessible

Teaching what is necessary for a particular set of tasks

e.g. teaching web-based tools like Chan Zuckerberg ID (formerly known as IDseq), along with the basics of file formats and skills for manipulating files

Training for specific pathogens

- Introducing existing workflows, tools, resources such as Galaxy workflows
- Experienced experts recommending most suitable options for the context
- Conveying best practices for sample selction/study design, data sharing, etc.

Developing bioinformatics skills

- Fundamental skills
- Intermediate
- Advanced

Highlighting, demonstrating best practices for programming, version control, project organisation

The Carpentries have excellent resources for training in concepts like

Example 1. Interpretation of genomic QC metrics

Refine sub- topics, objectives and learning outcomes

- Outline session objectives e.g.
  This session will cover topics in QC metrics, the process, thresholds, validating QC procedures and detecting contamination.

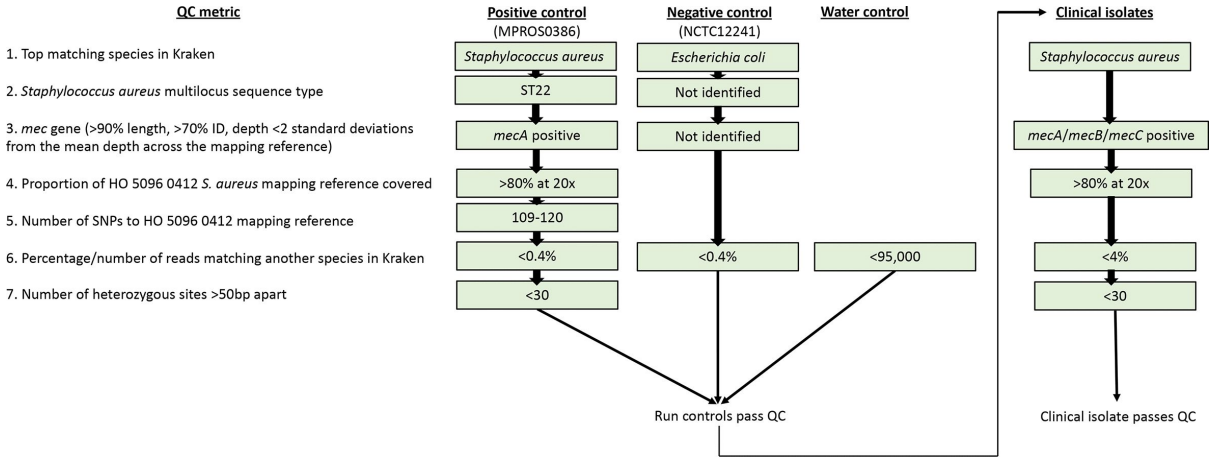
- Refine learning outcomes (dependant on target audience, goals, resources)
  E.g. For a mixed audience of laboratory scientists and bioinformaticians

  Example LOs - By the end of the session attendees should be able to
  - Describe QC metrics for genomic sequence data;
  - Apply relevant tools (or interpret information from a procedure) to detect contamination

Outcome-based strategies to teach genomic QC metrics

Expected (Learning) Outcome	Activities and assessment
Apply QC metrics and criteria to distinguish poor vs good quality genome sequences. Describe how QC thresholds are set; the type of controls used (QC thresholds may vary by microbial organism)	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. Ask learners to identify sample genomes with poor quality, they should describe their approach and outline their criteria/metrics for determination of what is poor or good quality.
Identify important checkpoints where errors may occur from sample processing to sequencing run.	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. What information (i.e. combination of various genomic QC metrics) helped to diagnose what went wrong in the upstream data collection, processing and/or sequencing steps?
Describe the impact of bad quality samples on interpretation	In groups provide learners with case studies on wrong interpretation, and wrong clinical/epidemiological actions that would have followed, caused by bad-quality samples; and discuss how interpretation changed once bad sample(s) were removed.
Describe the various sources of contamination (different species vs. strain contamination);	Selection of a diverse set of bad-quality samples. Use a quiz or poll determine whether contamination is due to different species vs strain cntamination

Resources and tools for application of QC metrics



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

Raven, K. E. *et al.* Defining metrics for whole-genome sequence analysis of MRSA in clinical practice. *Microbial Genomics* **6**, (2020).

Example 2. Data Interpretation - Phylogenetic trees

Refine sub- topics, objectives and learning outcomes

- Outline session objectives e.g.
  This session will cover topics basic concepts of phylogenetics and applications in genomic epidemiology to investigate microbial transmission

- Refine learning outcomes (dependant on target audience, goals, resources)
  E.g. For a laboratory scientist audience
  - Example LO - By the end of the session attendees should be able to
    - Read and interpret phylogenetic trees in the context of infectious diseases epidemiology.

Introduction to Phylogenetics concepts

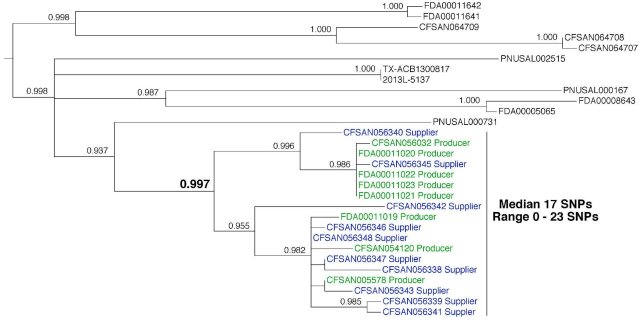
Introduction to Phylogenetic concepts

Present content (using slides or other means of introducing the topics)

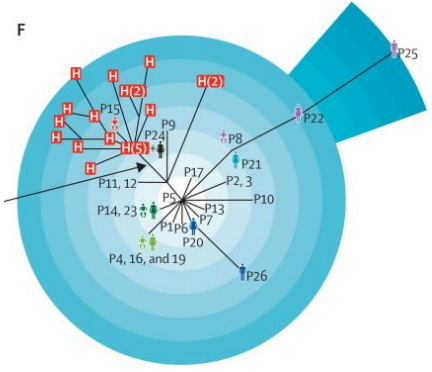
- Phylogenetics overview
- How are phylogenetic trees reconstructed from the number and pattern of shared mutations between strains (and assumptions)
- Definitions and nomenclature (“clade”, “tips”, “topology”, “branches”, etc.) and assumptions of phylogenetic tree reconstructions (e.g. clonal reproduction, mutation rates.)
- Epidemiology case studies
- Overview of online resources on how to interpret phylogenetics data

Using case studies to teach interpretation of phylogenetics

Example activities - Discuss using genomic epidemiology case-studies to demonstrate how to identify the source of foodborne outbreaks (e.g. *Salmonella* or *Listeria* outbreaks); hospital (e.g. MRSA or *Pseudomonas*); community (e.g. TB) or sexually transmitted outbreaks.



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



Phylogeny of the MRSA SCBU outbreak

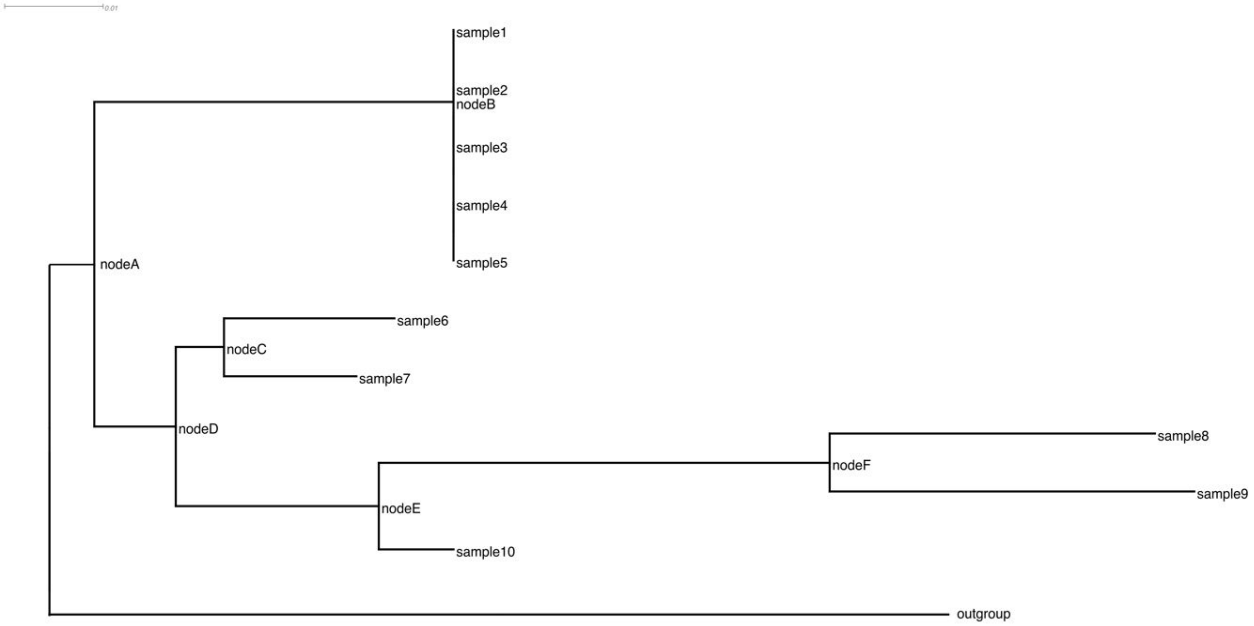
Pightling, A. W. *et al.* Interpreting Whole-Genome Sequence Analyses of Foodborne Bacteria for Regulatory Applications and Outbreak Investigations. *Frontiers in Microbiology* **9**, 1–13 (2018).

Example: MCQ activity - interpreting phylogenetic trees

- MCQ activity on how to interpret phylogenetic trees, with a particular focus on extracting strain relatedness information

- **Use Online resources** on how to read phylogenetic trees that introduce phylogenetic concepts and nomenclature including.
  - EBI Phylogenetics course - 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.

Activity: Practice in groups



What internal node corresponds to the most recent common ancestor of samples 8 and 10:

a. Node F
b. Node D
c. Sample 7
d. Node E

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

a. Samples 1 to 5
b. Samples 6 & 7
c. Samples 6 to 10
d. Samples 8 & 9

Review the phylogenetics activity in Groups

List in your groups

A.) What are the benefits to using this activity to teach phylogenetics?

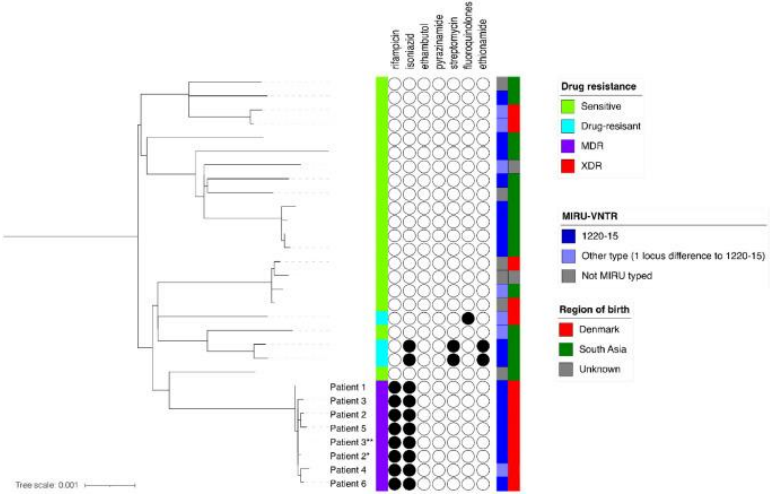
B.) What are the challenges in using such an activity?

C.) How would you improve on this activity?

Each group then reports back 1 point regarding the above questions

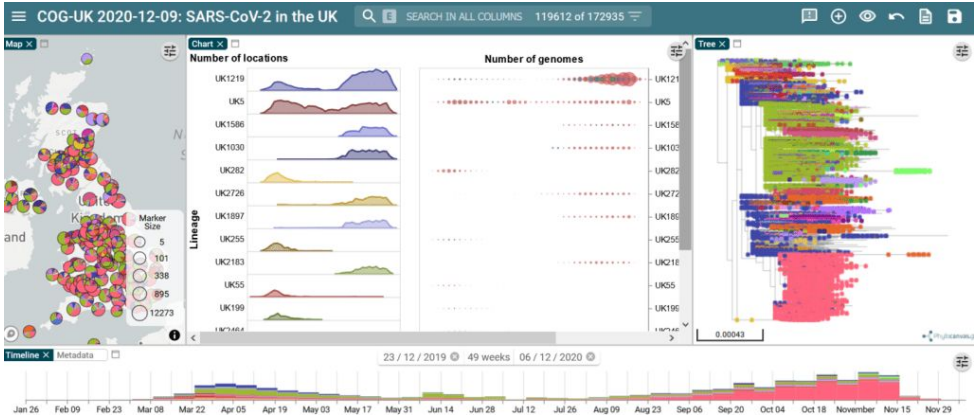
Example 3. Visualisation of genomic and epidemiological data

Sub-topic	Training approaches and resources
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.	Demonstrate how to use publicly-available tools such as Microreact or iTOL (https://itol.embl.de/) to visualise phylogenetic trees along with annotations of epidemiological data
Specialised tools such as Microreact (https://microreact.org/) or Nextstrain (https://nextstrain.org/) which include purpose-built functionality for genomic epidemiology investigations.	Demonstrate how to use of publicly-available tools such as Microreact or Nextstrain for interpreting phylogenetic trees and epidemiological data in genomic epidemiology investigations
Patient timeline plots - plots and data visualisation strategies used to visualise genomic and epidemiological data (e.g. patient timelines, networks, minimum spanning tree, blobograms, etc.)	Demonstrate how to interpret plots and data visualisation data used in real-world genomic epidemiology investigations. For example, timelines of patients visiting different hospital wards. Demonstrate how to identify common epidemiological markers (e.g. patients staying at the same hospital ward); transmission patterns such as super-spreaders based on the topology of the tree or shape of minimum spanning trees; etc.



iTOL

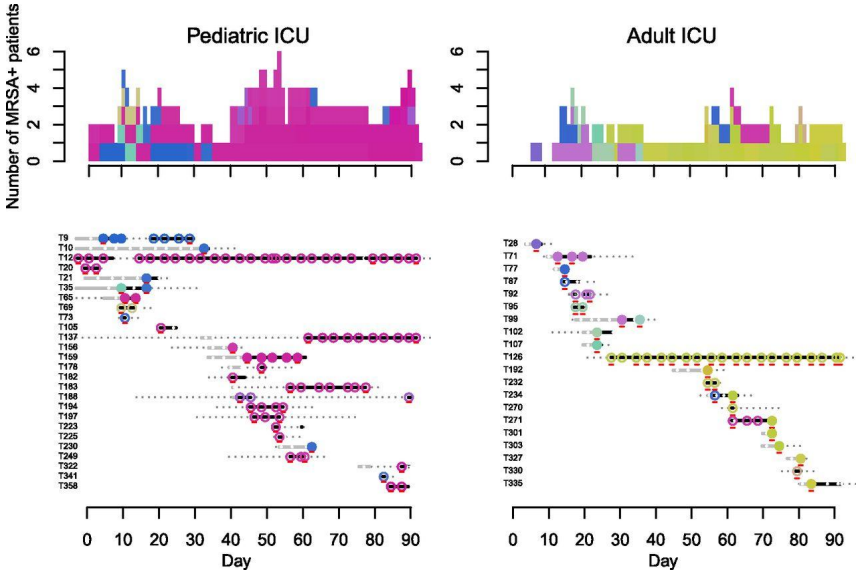
Phylogenetic tree and metadata of an MDR-TB outbreak strain in Denmark and contextual strains plotted using iTol

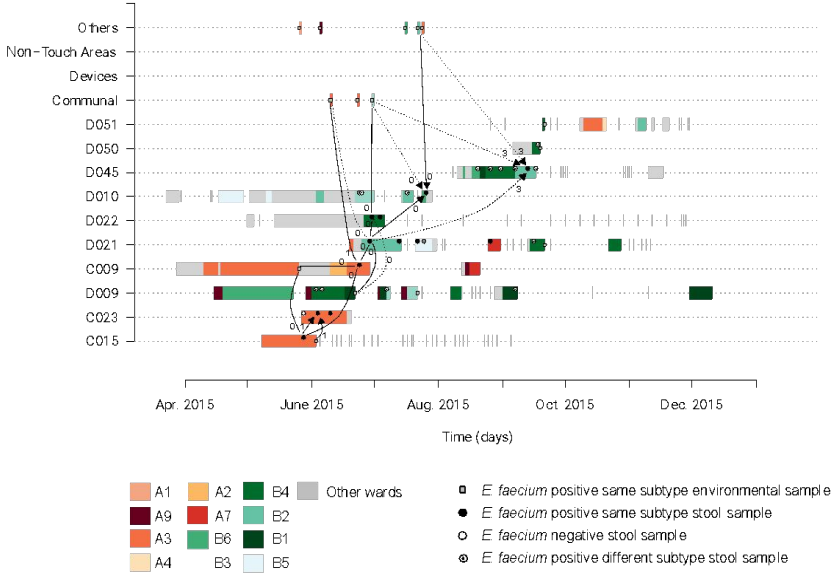


MicroReact

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

Timeline plots





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

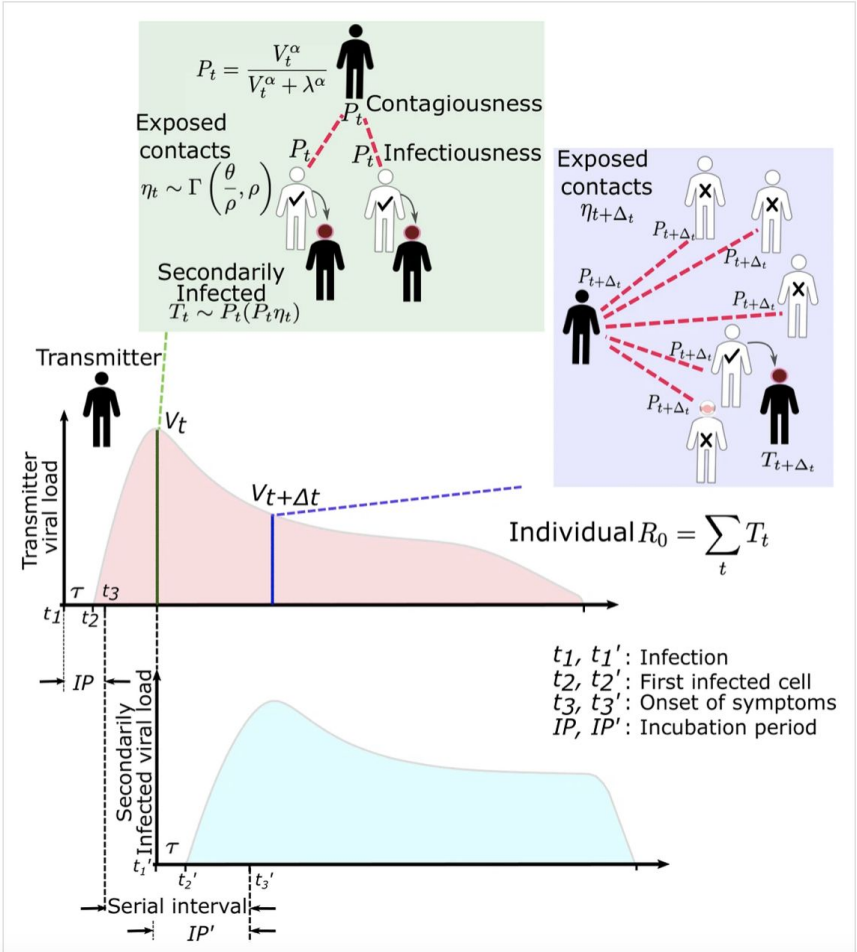
Group activity: Data Visualisation

Evaluate this figure and answer the following questions:

A.) Who is the intended audience for this figure?

B.) What would you require to train participants to make a figure such as this in terms of:

- Software?
- Computational resources?
- Input data?
- Trainer Expertise?



SARS-CoV-2 and influenza transmission model schematic.

https://elifesciences.org/articles/63537

Key considerations designing data analysis and interpretation training

- Using outcome based approaches is key to developing appropriate sub-topics for the target audiences
- Each sub-topic requires tailoring of tools and approaches for effective delivery of content
- Maximise the use of freely available tools, software and resources
- The learning environment will shape the complexity and depth as to what can be taught in bioinformatics
- Optimise *reproducibility* throughout to allow participants to integrate their own datasets later, and enable them to share with others