T3connect

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

Module 3C
Data Analysis and Integration
07/03/22

Silvia Argimón

wellcome connecting science

Centre for Genomic Pathogen Surveillance

BIG DATA INSTITUTE

UNIVERSITY OF OXFORD

Session Outline

Aim: How to communicate to trainees the importance of the integration of different sources of data for decision making

- (Genomic) surveillance objectives
- Data sources and challenges
- Data analysis and integration
- CGPS tools for data integration and interactive visualization
- Activity and discussion

wellcome connecting science

T3connect

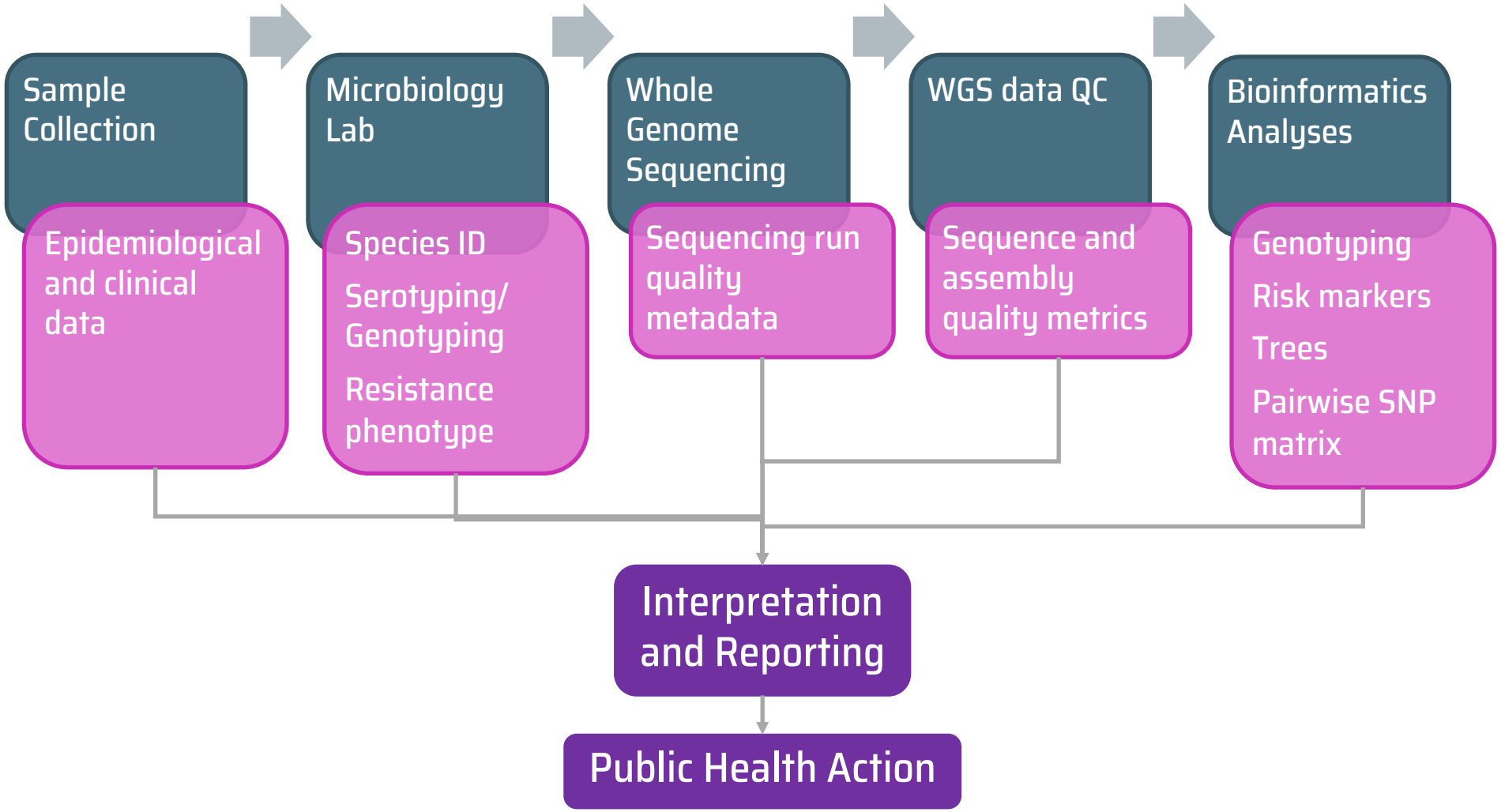
Centre for Genomic Pathogen Surveillance

Surveillance Objectives for Infection Control

- Identifying and mapping of high-risk clones (HiRiCs) and high-risk elements
- Understanding the routes of transmission and the role and behaviour of vectors in the dissemination of HiRiCs
- Understanding the role of different sources or reservoirs in the dissemination across environmental, animal, and human habitats

Hajo Grundmann (2014). Towards a global antibiotic resistance surveillance system: a primer for a roadmap, Upsala Journal of Medical Sciences, 119:2, 87-95

Genomic Surveillance - Data Sources and Integration



wellcome connecting science

T3connect

Centre for Genomic Pathogen Surveillance

Data Challenges 😱

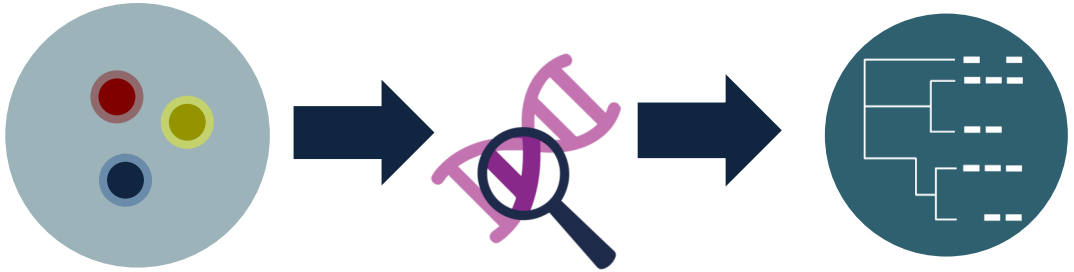
- Completeness and consistency
  - Collecting hospital doesn't always record the patient's diagnosis
  - Different labs might test different panels of antibiotics
- Interoperability
  - Different computer platforms
- Standardization
  - Vietnam vs Viet Nam, or 07/10/2019 vs 10/07/2019
  - CLSI, EuCAST, BSAC susceptibility breakpoints
- Different end-users
  - Physicians, infection control, hospital authorities, Ministry of Health
- Formats
  - Output of one tool is not in the required format to use as input of the next tool
  - Integrating data from different sources (.pdf, .xlsx, .csv, .tre)

wellcome connecting science

T3connect

Centre for Genomic Pathogen Surveillance

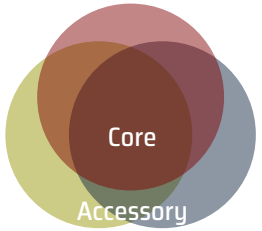
Genomic Surveillance - Data analysis



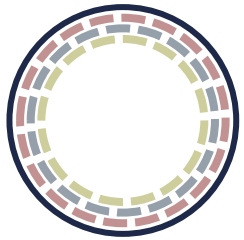
Extracting epidemiologically relevant information from the genomes

- Genetic relatedness
  - Genotyping (MLST, pathogen-specific)
  - Phylogenetic relationships (tree)
- Presence of markers of risk

Pangenome Analysis



Mapping to a Reference

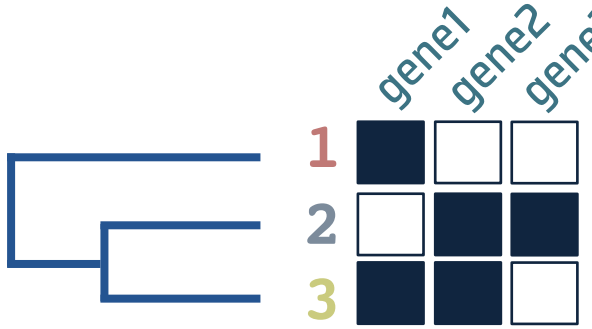


BLAST assemblies

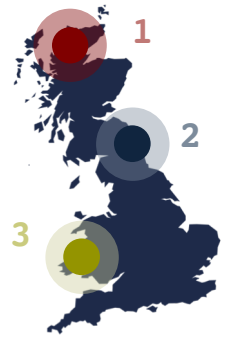
map reads

```
1  AATCGCTTTACGACCAG...
2  AATGGCTTTATGACAAG...
3  AATCGCTTTATGACAAG...
      *       *   *
```

SNP-based tree

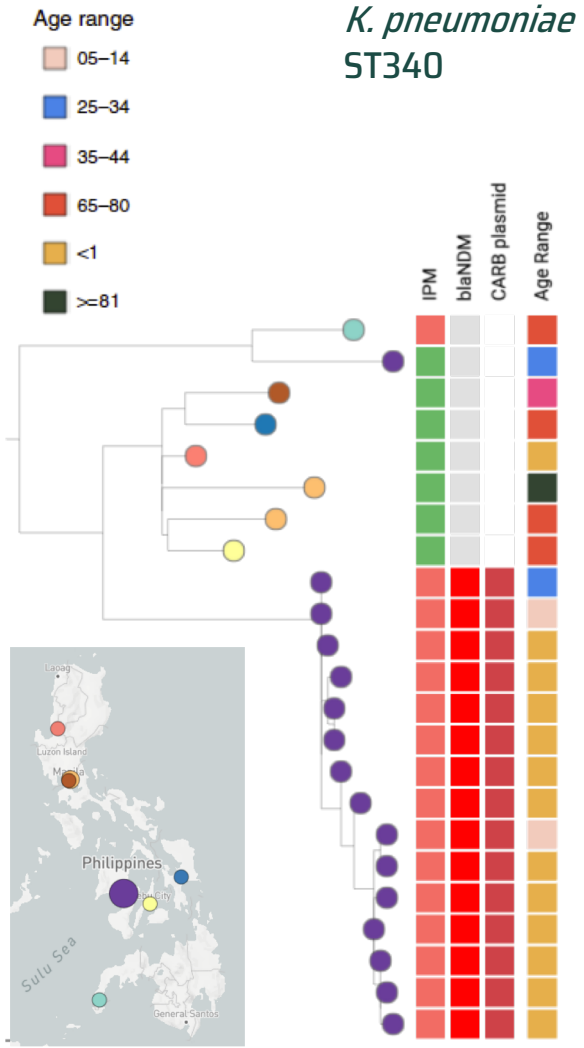


Newick format: ((2,3)1);



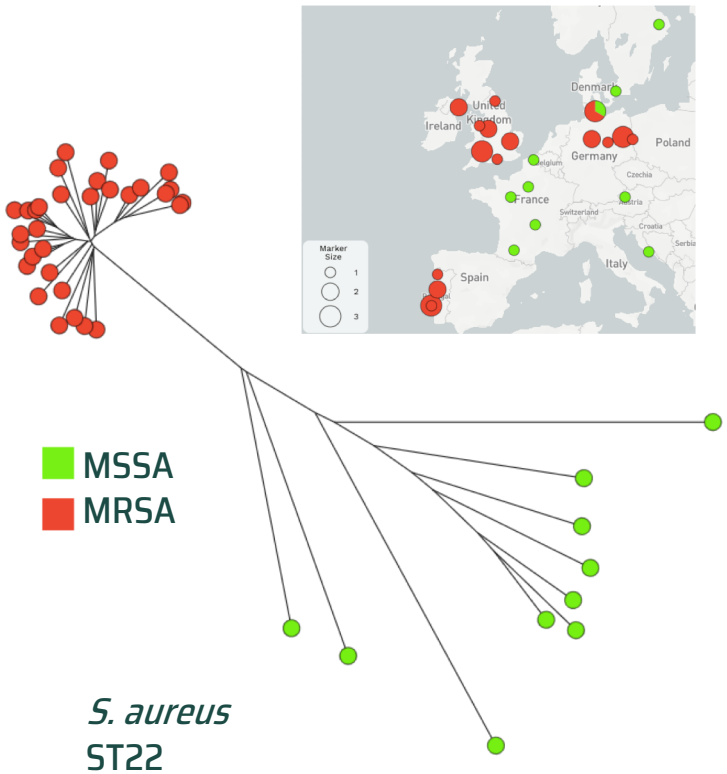
For examples of analytic tools see Module 1B - Microbial genomics resources

High-risk clones and genomic epidemiology



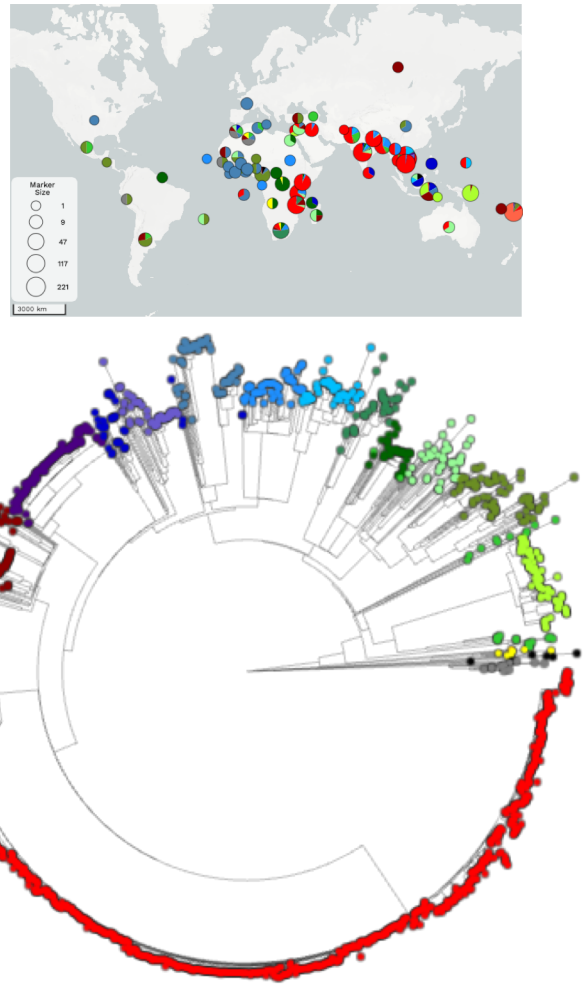
K. pneumoniae
ST340

- Clonal Relatedness
- Abundance
- Spatial/temporal distribution
- Risk properties



S. aureus
ST22

S. Typhi
4.3.1 (H58)



Argimón S. et al. 2020. Nat Comms 11:2719

Aanensen D. et al. 2016. Mbio 7(3):e00444-16

Wong V. et al. 2015. Nat Genet 47:632

CGPS - Free Web Applications for Pathogen Surveillance

Collection

epicollect

https://five.epicollect.net

Integration

Data-flo

https://data-flo.io/

Interactive Visualization

Microreact

http://microreact.org

Analysis

Pathogenwatch

https://pathogen.watch

@Pathogenwatch @MyMicroreact @EpiCollect

wellcome connecting science

T3connect

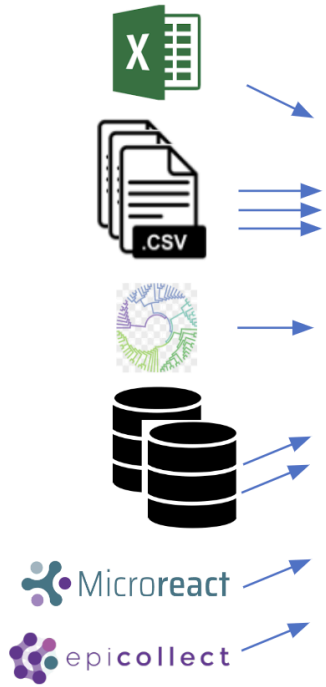
Centre for Genomic Pathogen Surveillance

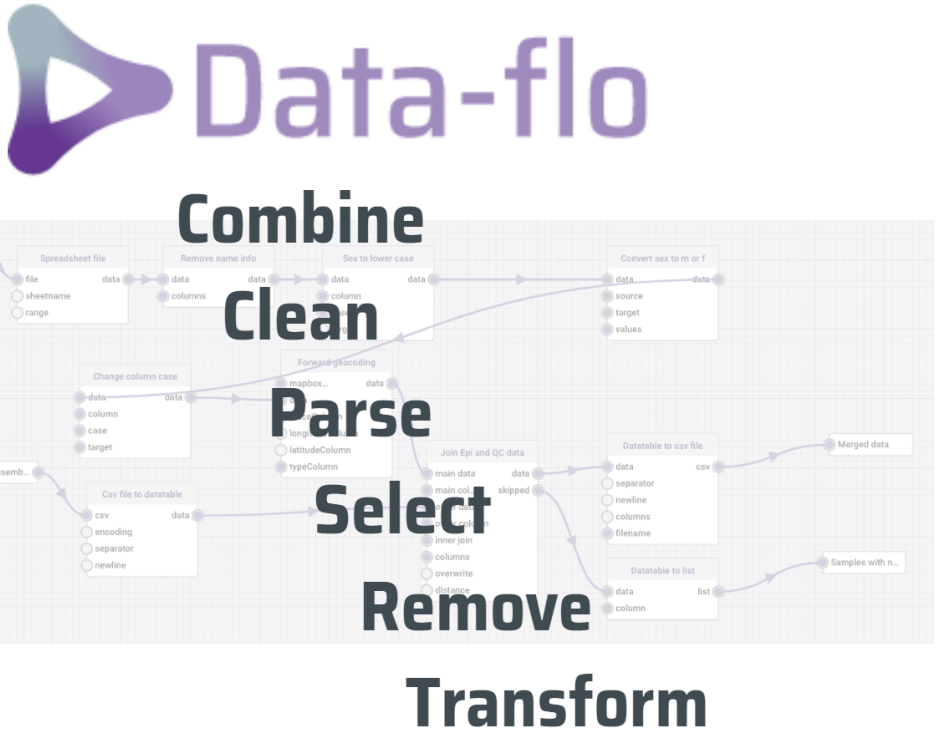
Data-flo

https://data-flo.io/

Customised integration and manipulation of diverse data via a simple drag and drop interface

SOURCES





DESTINATIONS

- Microreact
- Google Drive
- CSV
- TSV
- SQLite DB file
- Etc etc

wellcome connecting science

T3connect

Centre for Genomic Pathogen Surveillance

Microreact

http://microreact.org

Interactive visualization of clustering (trees), geographic (map) and temporal (timeline) data.

Whole genome sequencing reveals potential spread of *Clostridium difficile* between humans and farm animals in the Netherlands, 2002 to 2011

C W Knetsch[1], T R Connor[2], A Mutreja[3], S M van Dorp[1], I M Sanders[1], H P Browne[3], D Harris[3], L Lipman[4], E C Keessen[4], J Corver (j.corver@lumc.nl)[4], E J Kuijper[4], T D Lawley[3]

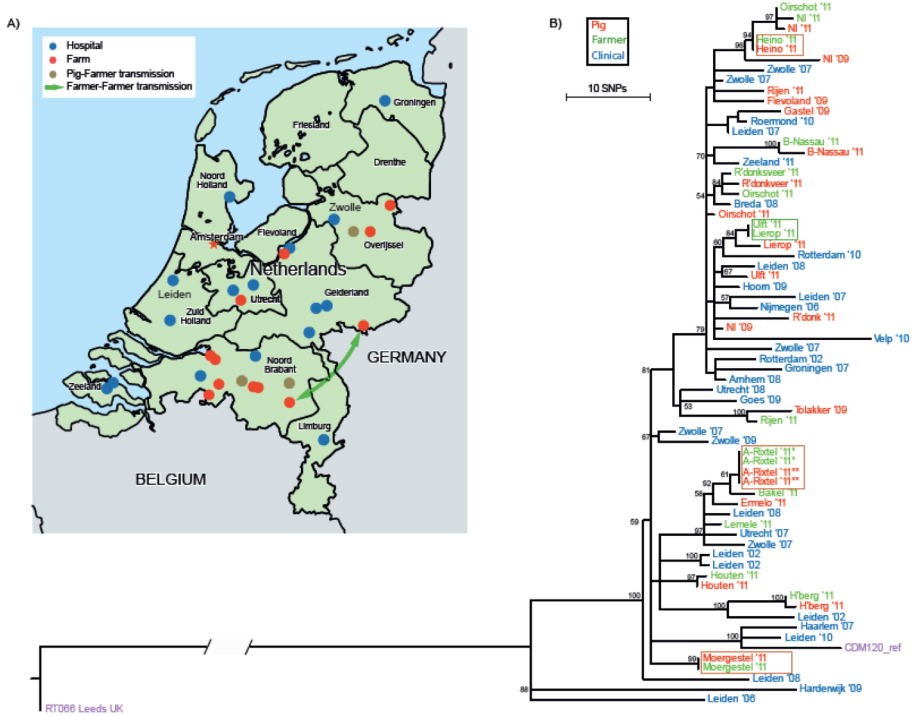


TABLE 1A

Clostridium difficile type 078 isolates used in this study, the Netherlands, 2002–11 (n=65)

R_L#T[a]	Year	City	RT	Isolate	Source	Related isolates	Association	ENA ID[b]
8080_2#24	2006	Leiden	078	6072310	Clinic	Non-outbreak	Healthcare	ERS138026
8080_2#25	2006	Nijmegen	078	6086336	Clinic	Non-outbreak	Healthcare	ERS138027
8080_2#26	2007	Leiden	078	7001233	Clinic	Non-outbreak	Healthcare	ERS138028
8080_2#27	2007	Groningen	078	7004578	Clinic	Non-outbreak	Unknown	ERS138029
8080_2#28	2007	Utrecht	078	7005405	Clinic	Non-outbreak	Unknown	ERS138030
8080_2#29	2007	Zwolle	078	7021455	Clinic	Non-outbreak	Healthcare	ERS138031
8080_2#30	2007	Zwolle	078	7044912	Clinic	Non-outbreak	Community	ERS138032
8080_2#31	2007	Zwolle	078	7066827	Clinic	Non-outbreak	Community	ERS138033
8080_2#32	2007	Zwolle	078	7071308	Clinic	Non-outbreak	Healthcare	ERS138034

FIGURE 2

Phylogenetic cluster showing relatedness of *Clostridium difficile* clinical, pig and farmer isolates, the Netherlands, 2008–11 (n=4)

TABLE 3

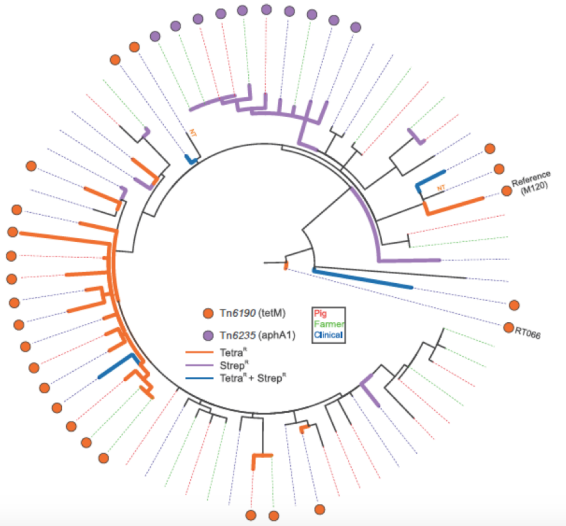
Results of Antimicrobial susceptibility testing

Isolate	Source	Strepto-mycin	Tn*6235*	Tetra-cycline	Tn*6190*
6072310	Clinic		Absent		Present
6086336	Clinic		Absent		Present
7001233	Clinic	NT	Absent		Absent
7004578	Clinic		Absent		Present
7005405	Clinic		Present		Absent
7021455	Clinic		Absent		Present
7044912	Clinic		Present		Absent
7066827	Clinic		Absent		Present
7071308	Clinic		Absent		Absent
7086074	Clinic		Absent		Absent
7091952	Clinic		Absent		Present
8011061	Clinic		Absent		Absent
8013820	Clinic		Absent		Absent
8051728	Clinic		Absent		Present
8055344	Clinic		Present		Absent

Isolate	Source	Strepto-mycin	Tn*6235*	Tetra-cycline	Tn*6190*
129280	Clinic		Absent		Absent
H205	Farmer		Present		Absent
B37.3	Pig		Present		Absent
53737	Clinic	NT	Absent	NT	Present
47337	Clinic		Absent		Present
H102	Farmer		Absent		Absent
B31.3	Pig		Absent		Absent
B17.3	Pig		Absent		Absent
H121	Farmer		Absent		Present
B27.7	Pig		Absent		Absent
H230	Farmer		Absent		Present
H189	Farmer		Absent		Absent
B23.6	Pig		Absent		Absent
H205	Farmer		Present		Absent
B15.1	Pig		Absent		Absent

FIGURE 3

Phylogeny of *Clostridium difficile* 078 isolates showing the presence of antimicrobial resistance determinants, the Netherlands, 2002–11 (n=65*)



wellcome connecting science

T3connect

Centre for Genomic Pathogen Surveillance

Basic concepts in data analysis and integration

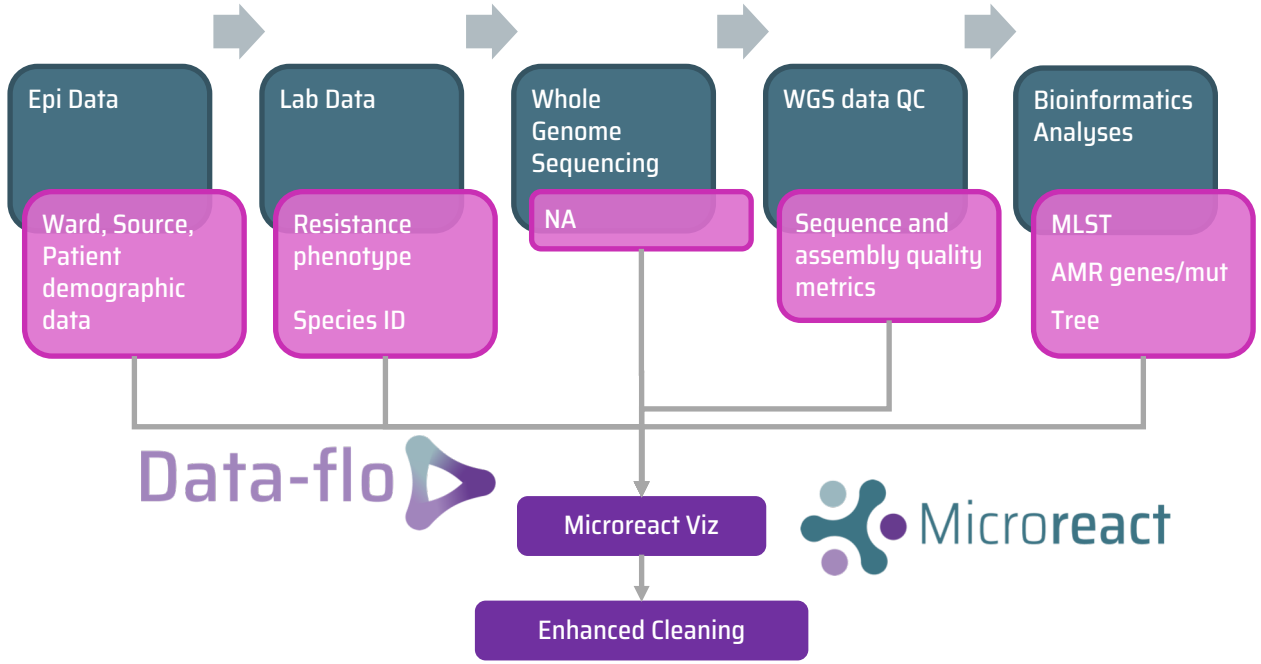
Domain Concepts	Strategies	Assessments	Resources
Specific analytic tool/pipeline	Demo/tutorial with dummy data. Presentation with link to tools/resources	Exercise with “real data”	Module 1B https://lms.wellcomeconnectingscience.org/mod/forum/view.php?id=2525
Different types of data for genomic surveillance	Presentation Poll Group Activity	Wrap-up discussion	https://docs.data-flo.io/using-data-flo/data/data-types
Importance of data harmonization	Presentation (Bonus) Group Activity	Wrap-up discussion	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8847733/
Data integration and visualization	Presentation Poll Group Activity	Wrap-up discussion	data-flo https://docs.data-flo.io/introduction/readme Microreact https://docs.microreact.org/ Tidyverse https://github.com/rstudio/cheatsheets/blob/main/data-transformation.pdf Nextstrain https://nextstrain.org/ Phandango https://jameshadfield.github.io/phandango/#/
Genomic surveillance/epidemiology	Presentation (Examples) Group Activity	Wrap-up discussion Interpretation of case study	https://alliblk.github.io/genepi-book/ https://www.futurelearn.com/courses/pathogen-genomics-a-new-era-in-global-health-surveillance-and-strategy https://www.futurelearn.com/courses/genomics-covid-19

Module 3C Exercise

Data analysis and integration for decision making

Hospital outbreak investigation → Enhanced cleaning

- Work in pairs
- Read the Background
- Activity 1. Data sources
- Activity 2. Integration and Viz
- Activity 3. Reflection (table groups)
- Session wrap-up.



wellcome connecting science

T3connect

Centre for Genomic Pathogen Surveillance

Module 3C Wrap-up

Aim: Highlight the importance of the integration of different sources of surveillance data for decision making

Presentation (passive learning)

Activities (active learning)

Which domain concepts covered in this session (or similar ones) would you foresee having to teach in the future? How would you do it? Note: See presentation slides 2 and 11 for reference to domain concepts.

What do you think are the pros and cons of using interactive web tools such as data-flo and Microreact for teaching?

Can you think of alternative ways to teach this module?

wellcome connecting science

T3connect

Centre for Genomic Pathogen Surveillance

References

Hajo Grundmann. 2014. Towards a global antibiotic resistance surveillance system: a primer for a roadmap. Upsala Journal of Medical Sciences, 119:2, 87-95 doi:10.3109/03009734.2014.904458

Argimón S, Abudahab K, Goater R, Fedosejev A, Bhai J, Glasner C, Feil E, Holden M, Yeats C, Grundmann H, Spratt B, Aanensen D. 2016. Microreact: visualizing and sharing data for genomic epidemiology and phylogeography. M Gen, 2(11): doi:10.1099/mgen.0.000093

Knetsch CW, Connor TR, Mutreja A, van Dorp SM, Sanders IM, Browne HP, Harris D, Lipman L, Keessen EC, Corver J, Kuijper EJ, Lawley TD. 2014. Whole genome sequencing reveals potential spread of *Clostridium difficile* between humans and farm animals in the Netherlands, 2002 to 2011. Euro Surveill, 19(45):20954. doi: 10.2807/1560-7917.es2014.19.45.20954

wellcome connecting science

T3connect

Centre for Genomic Pathogen Surveillance

Acknowledgements

This course was developed by a collaboration between the Centre for Genomic Pathogen Surveillance and Wellcome Connecting Science. It was brought to you by *T3Connect – Data Science and Genomic Pathogen Surveillance Training Programme*, funded by *UKRI*.

This module contains materials from the following sources:

- Storyset | Customize, animate and download illustration for free

wellcome connecting science

T3connect

Centre for Genomic Pathogen Surveillance

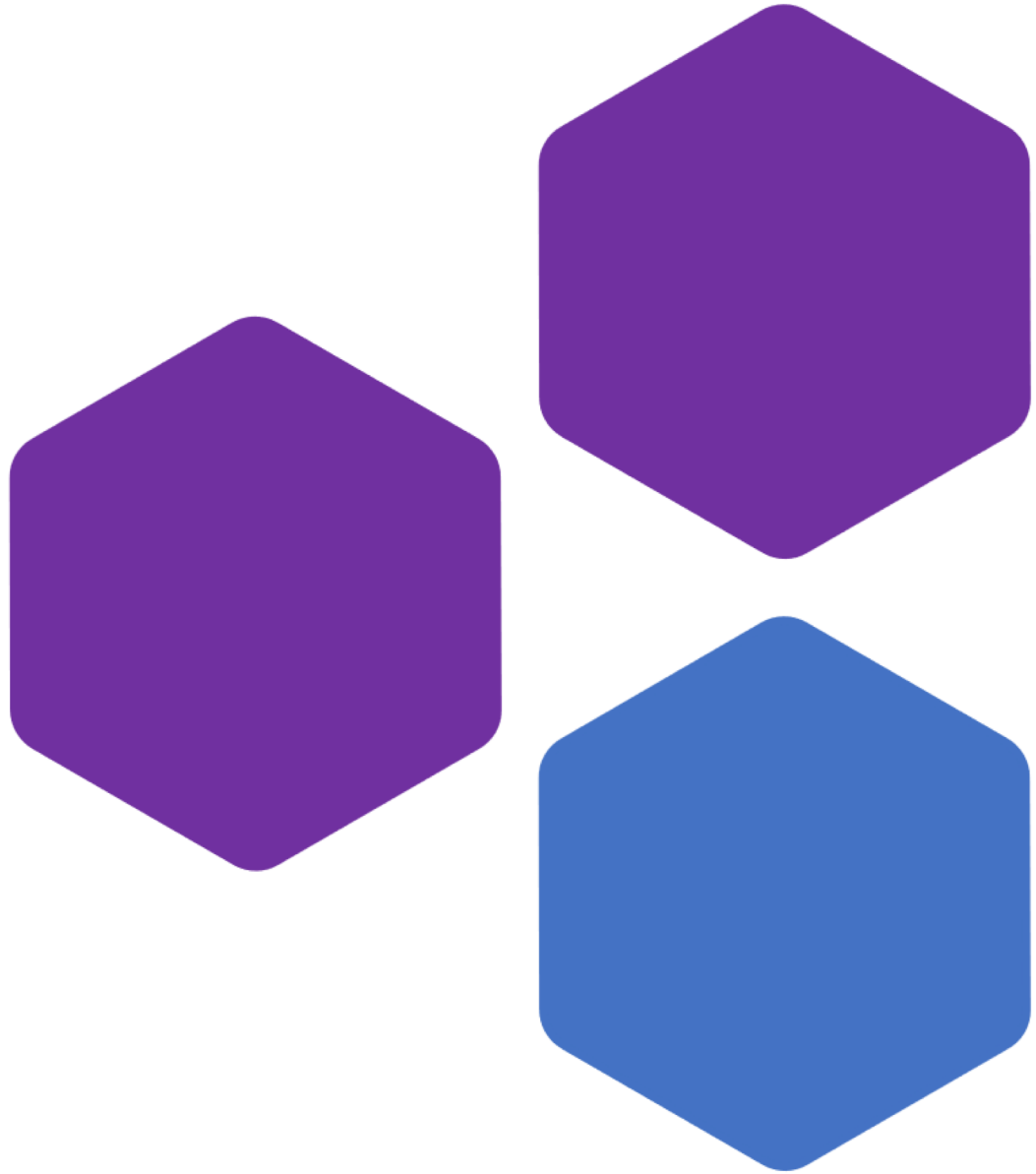
Creative Commons

wellcome connecting science

T3connect

Centre for Genomic Pathogen Surveillance

Thank you



T3connect