Empowering AMR Research through R: Analysis and Visualisation

Session 1: AMR Data Visualisation in R

Niamh Lacy-Roberts, Global Capacity Building Group, DTU, nlac@food.dtu.dk

18th June 2025

Welcome to the virtual training on "Empowering AMR Research through R: Analysis and Visualisation"!

Aims:

- In this hands-on training, we'll explore how to apply R to antimicrobial resistance (AMR) and whole genome sequencing (WGS) data using real-world examples.
- You'll learn how to create publication-ready plots, heatmaps, and phylogenetic visualisations that support effective interpretation and communication of AMR trends.
- Whether you're working with national surveillance data or outbreak investigations, this course aims to provide you with the practical tools and confidence to explore, visualise, and share AMR insights using R.

By the end of this workshop, participants will be able to:

- Create advanced AMR visualizations using ggplot2 and patchwork.
- Summarise and visualise resistance data by location, organism, and time.
- Generate clustered heatmaps using phenotypic and genotypic resistance data.
- Integrate WGS-derived resistance gene data into AMR analyses.
- Visualise phylogenetic trees using ggtree() and annotate them with AMR data using gheatmap().
- Interpret and communicate trends in AMR using effective plots.

Schedule and agenda

This virtual training takes place over two sessions that will be held on Wednesday, June 18 and June 25 from 9:00 to 12:30 (CEST).

Session 1: AMR Data Visualisation in R (18 June, today! ☺)

- AMR Data Wrangling and Visualisation
- Advanced Plotting Techniques
- Heatmap Visualisation
- Practical Exercise Briefing and Q&A

The practical exercises presented in Session 1 should be completed between sessions. The answers will be provided at the beginning of Session 2.

Session 2: Phylogeny and AMR Integration in R (25 June)

- Practical Exercise Review and Recap
- Phylogenetic Tree Basics in R
- Annotating Trees with AMR Data
- Wrap-Up and Q&A

Required software

- Hopefully you have installed and ready to go on your computer:
  - R version: 4.4.2 or later
  - RStudio version: 2024.12.0+467 or another recent stable release

R libraries we will use today

- tidyverse, pheatmap, ggtree, janitor, ComplexHeatmap, circlize

Let’s quickly recap, what is AMR?

- Antimicrobial Resistance (AMR) occurs when bacteria, viruses, fungi, and parasites evolve and stop responding to medicines, making infections harder to treat.
- It’s a global health threat: In 2019, AMR was directly responsible for over 1.27 million deaths globally — more than HIV or malaria.
- Drivers include overuse of antibiotics in humans and animals, poor sanitation, and lack of access to diagnostics and surveillance.

What is AMR continued

- Many AMR genes are located on plasmids, which are mobile genetic elements.
- This means AMR genes can transfer between different bacteria, accelerating the spread of resistance, even across species.
- WGS helps detect both chromosomal and plasmid-borne resistance genes.

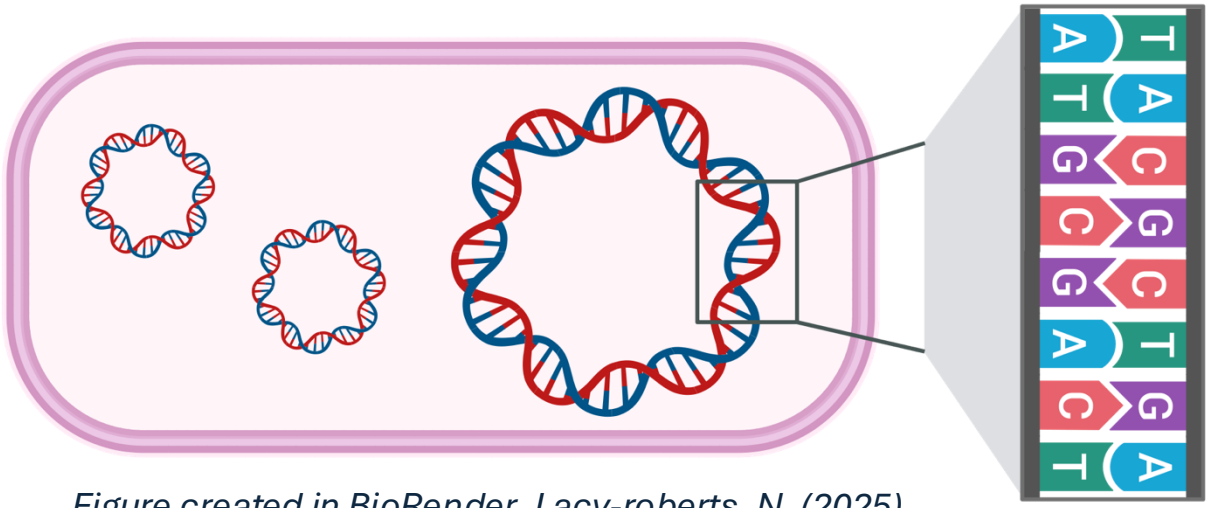


Figure created in BioRender. Lacy-roberts, N. (2025) https://BioRender.com/h1tar74.

And what about WGS?

- Whole Genome Sequencing (WGS) determines the complete DNA sequence of an organism's genome in a single run.
- In AMR surveillance, WGS is used to:
  - Identify bacterial species
  - Predict resistance genes (genotype)
  - Track outbreaks through phylogenetics
  - Understand transmission dynamics
  - And so on…!

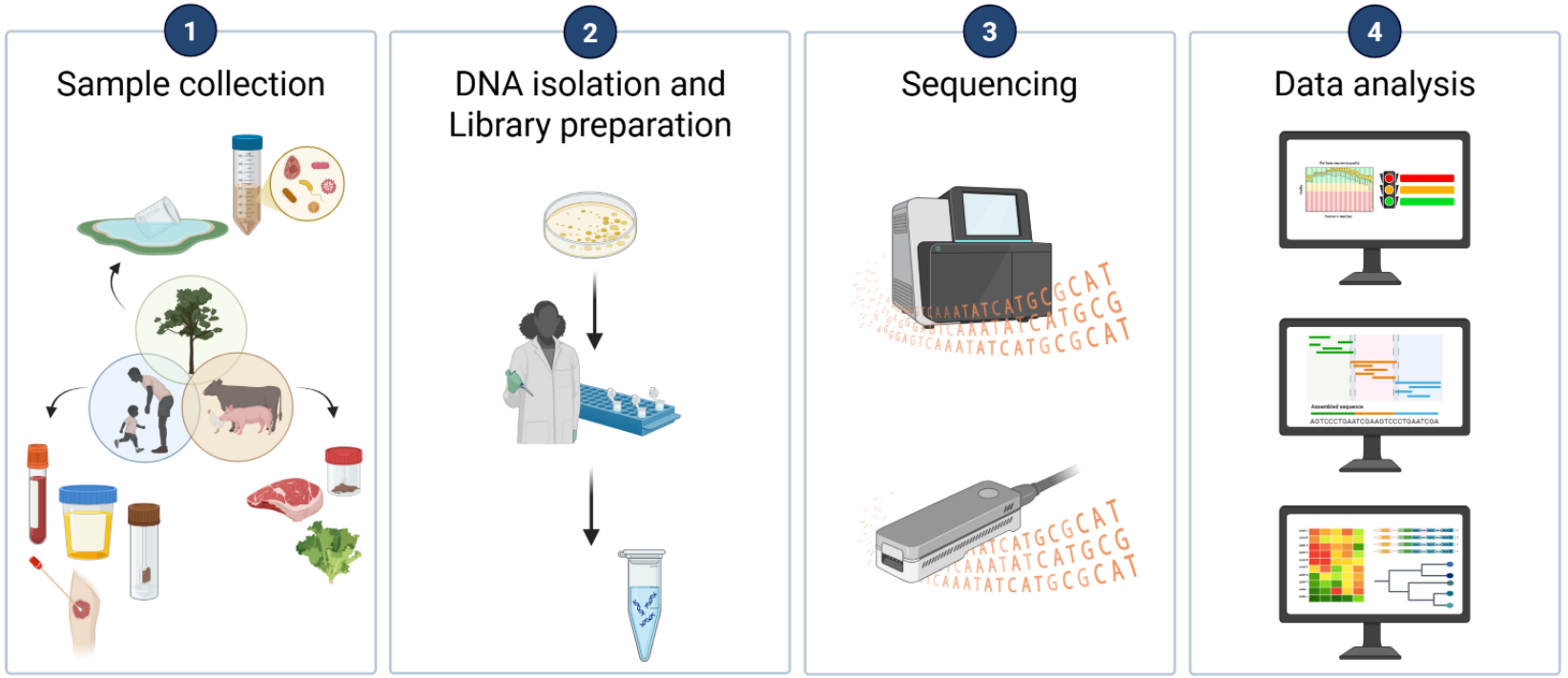


Figure created in BioRender. Nilsson, P. (2025) https://BioRender.com/lqzgrel

Phenotype vs. Genotype in AMR

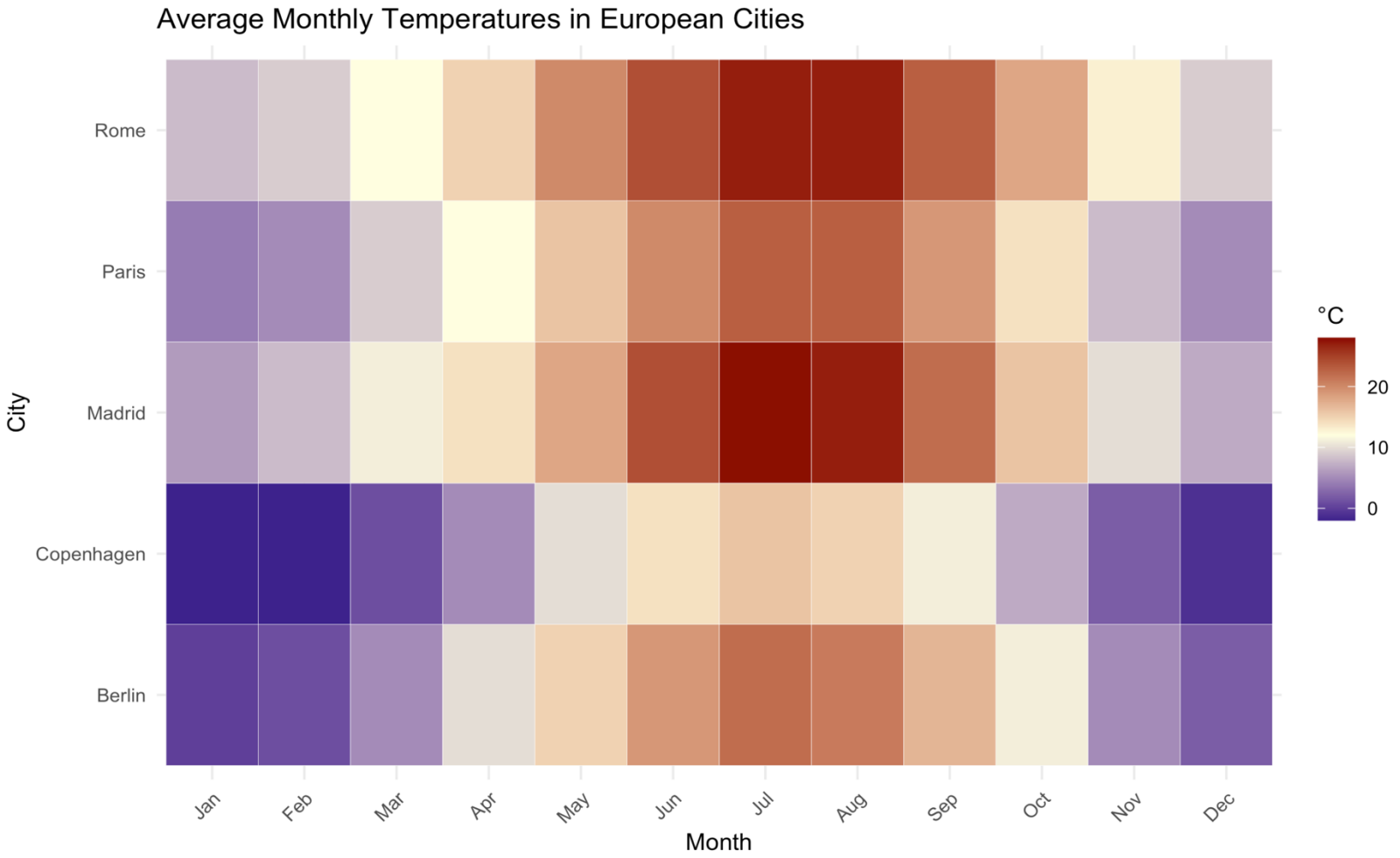
- Phenotypic resistance = lab testing (e.g., disk diffusion, MIC)
- Genotypic resistance = prediction based on resistance genes from WGS
- WGS allows us to predict AMR, compare across locations and years, and even link data to outbreaks.

- Linking AMR phenotype + genotype + context (e.g., time, country) allows for:
  - Better surveillance
  - Early outbreak detection
  - Informed public health action

What is a Heatmap?

- One of the main aims of this course is that you'll learn how to create heatmaps in R, but what is a heatmap?
- A heatmap is a data visualization that uses color to represent values in a matrix or table.
- Often used to show patterns, relationships, or intensities across multiple variables at once.

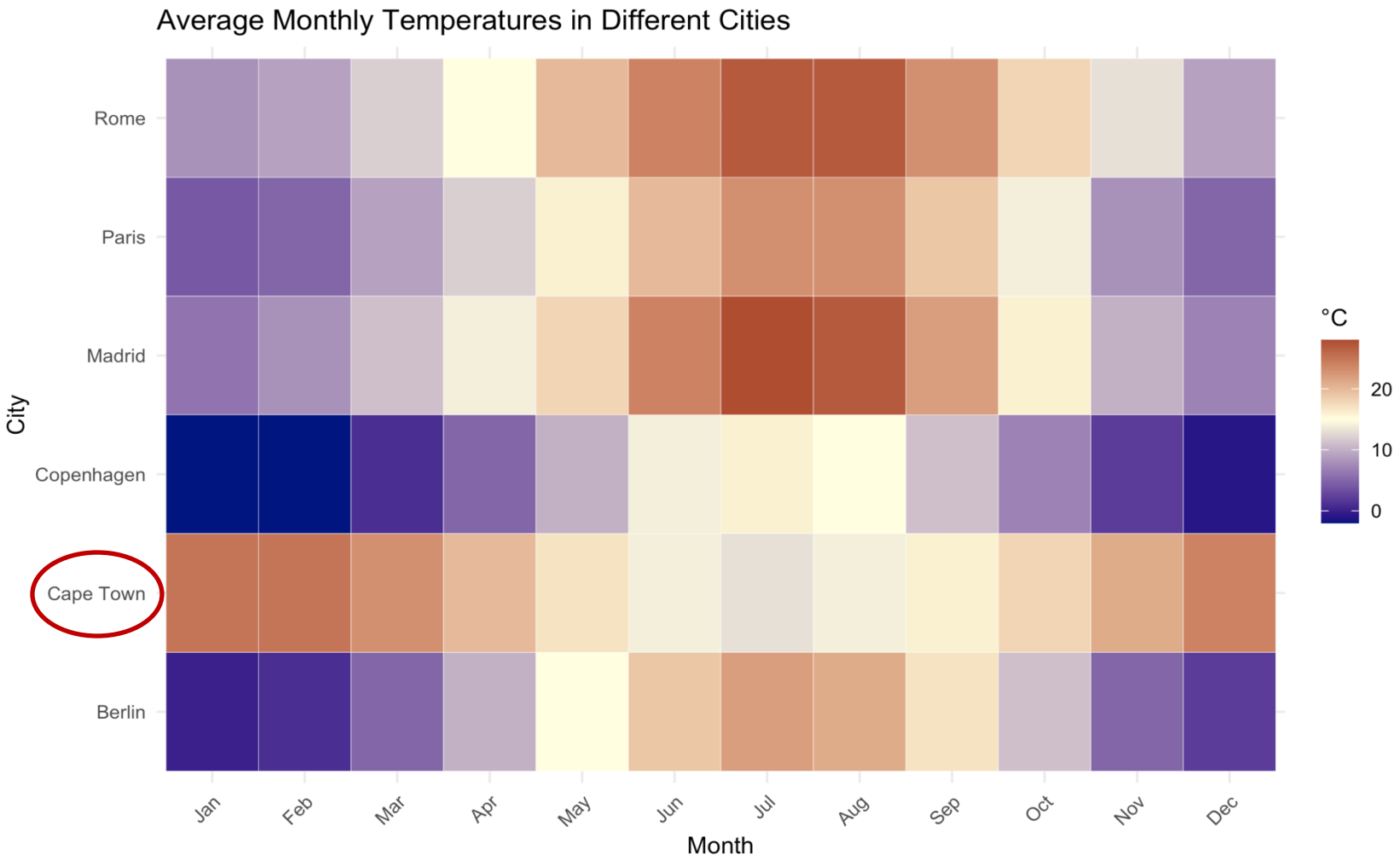
Heatmap example, European cities monthly temperature



What this heatmap shows:

- Each row is a city.
- Each column is a month.
- Colors represent average monthly temperatures (°C).
- Cooler colors = colder temperatures (e.g., winter)
- Warmer colors = hotter temperatures (e.g., summer).

Heatmap example, European cities + Cape Town monthly temperature



- Notice how cities in the Northern Hemisphere are cold in Jan–Feb and warm in Jul–Aug.
- Cape Town is in the Southern Hemisphere.
- It has summer when Europe has winter, and vice versa.
- This illustrates how heatmaps can visually reveal "opposite" seasonal trends and geographic differences.

Heatmaps for AMR

- In AMR and genomics, heatmaps are commonly used to show:
  - Presence/absence of resistance genes
  - Levels of resistance across antibiotics
  - Sample similarities, especially when combined with clustering

Example: Heatmap of AMR Profiles from a Surveillance Study

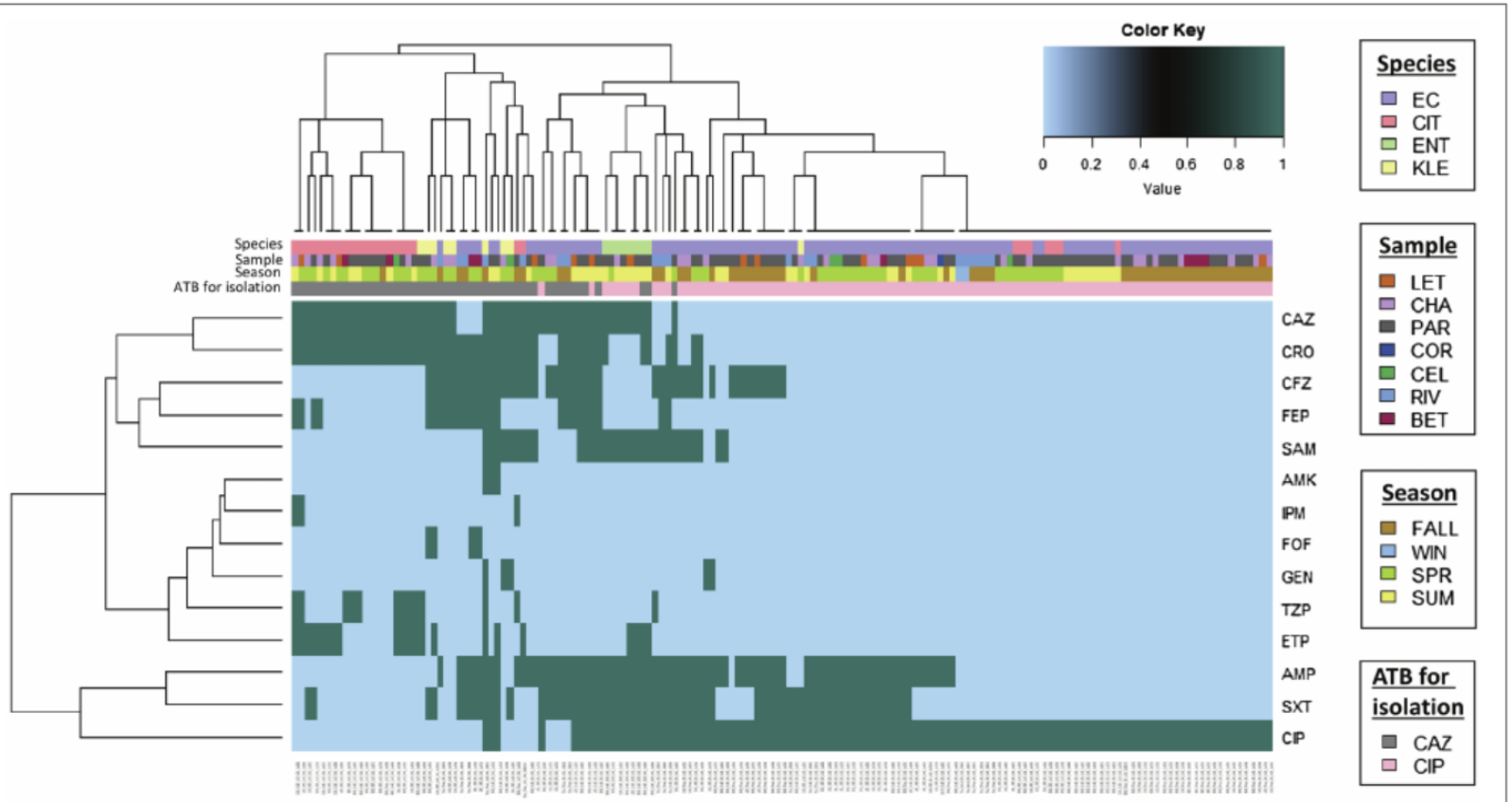


FIGURE 2 | Heat map representation of antimicrobial resistance (AMR) profiles. Cluster analysis of AMR profiles, in the vertical axis the abbreviations for the antimicrobial tested and in the horizontal axis are the isolates obtained in this study. Light blue represents sensitivity and dark green represents resistance. Five clusters were found and are labeled from 1 to 5 in the dendrogram.

- This heatmap shows phenotypic AMR profiles from Enterobacterales isolated from vegetables and river water in Chile.
- Rows represent antibiotics tested.
- Columns represent bacterial isolates.
- Dark green indicates resistance, and light blue indicates sensitivity.
- A hierarchical cluster analysis groups isolates (top) and antibiotics (left) based on similar resistance patterns.
- The colored bars above the heatmap show isolate metadata:
  - Species
  - Sample source (e.g., river water, vegetables)
  - Season of isolation
  - Antibiotic used for selection

Díaz-Gavidia, C., Barría, C., Rivas, L., García, P., Alvarez, F. P., González-Rocha, G., ... & Moreno-Switt, A. I. (2021). Isolation of ciprofloxacin and ceftazidime-resistant Enterobacterales from vegetables and river water is strongly associated with the season and the sample type. *Frontiers in microbiology*, *12*, 604567. https://www.frontiersin.org/journals/microbiology/articles/10.3389/fmicb.2021.604567/full

A Quick Refresher: tidyverse

- A collection of R packages designed for data wrangling and visualization
- Built on consistent grammar and piping (%>% or |>)
- Core packages:
  - dplyr for data manipulation
  - ggplot2 for visualization
  - And many more

Handy Helper Packages: janitor and stringr

janitor

- Designed to clean messy data quickly.
- clean_names() makes column names consistent:
  - Converts to lowercase
  - Replaces spaces and punctuation with underscores
  - Makes names unique
- Great for prepping data from Excel or CSV files.

stringr

- Part of the tidyverse, simplifies string manipulation.
- Examples:
  - str_to_title() – Capitalizes first letter of each word (e.g. "denmark" → "Denmark")
  - str_to_upper() – Converts strings to uppercase (e.g. "r" → "R")

Consistent syntax and good integration with dplyr.

Wide vs. Long Data Format (tidyr)

Wide format:

- Each variable gets its own column
- One row per subject/sample
- Common in spreadsheets and lab data

Isolate	Ciprofloxacin	Ampicillin	Colistin
ISO_001	R	S	R
ISO_002	S	R	S

You can use pivot_longer() from the tidyr package to go from wide → long

Long format:

- One column for variable name (e.g. antibiotic)
- One column for value (e.g. resistance)
- Better for plotting and analysis in R

Isolate	Antibiotic	Resistance
ISO_001	Ciprofloxacin	R
ISO_001	Ampicillin	S
ISO_001	Colistin	R
ISO_002	Ciprofloxacin	S
ISO_002	Ampicillin	R
ISO_002	Colistin	S

Comparing **pheatmap** and **ComplexHeatmap**

- The pheatmap package is great for creating quick and simple heatmaps with basic annotations.
- It's easy to use and works well for exploratory analysis or training.
  - However, it has limited flexibility when it comes to layering complex metadata or customizing layout.

- In contrast, ComplexHeatmap is highly customizable and designed for more advanced or publication-quality figures.
- It supports multiple annotations, split heatmaps, and integration with other plots.
- While it has a steeper learning curve and more verbose syntax, it offers much greater control over appearance and layout.
- It's especially useful when working with large datasets or when you need to combine multiple heatmaps and annotations in one figure.

Datasets overview

During this session, we'll work with two primary datasets that simulate AMR surveillance data across different countries, years, and bacterial species. These datasets are fictional but modeled on real-world formats to help you practice common data cleaning and visualization tasks.

- Dataset 1: amr_phenotype_eu.csv
- Purpose: This smaller dataset is ideal for quick prototyping and learning basic data wrangling and visualization techniques.
- Contents:
  - Isolate metadata: isolate_id, species, country, host, source, year
  - Phenotypic resistance results to 4 antibiotics: ciprofloxacin, ampicillin, tetracycline, colistin

- Dataset 2: amr_phenotype_larger.csv
- Purpose: A larger dataset used for more complex visualizations and grouped summaries, such as multi-faceted plots or resistance proportions over time.
- Contents:
  - Similar structure to the smaller dataset but includes 300 isolates from multiple countries, species, and years.

- Take-Home Dataset: amr_phenotypes_dirty.csv

Take-Home Exercise: Cleaning and Visualizing AMR Phenotypes with Heatmaps

Your Task:

- Download the "dirty" dataset.
- Data Cleaning:
  - Clean the species column to harmonize values like "E. Coli", "e.coli", "Salmonella sp.", etc.
  - Handle missing or empty resistance values ("", NA), and recode:
    - "R" → 1
    - "S" → 0
    - Blank or missing → leave as NA or impute sensibly.

Matrix Preparation:

- Create a resistance matrix with rows = isolate_id, columns = antibiotics.
- Make sure row names are unique.

Heatmap Visualization:

- Create a basic heatmap with no clustering or annotations.
- Then add annotations for:
  - species (color-coded)
  - country (color-coded)
- Optionally, arrange the heatmap by country or species.

Thank you for listening ☺

Acknowledgements

The creation of this training material was commissioned by ECDC to the Technical University of Denmark with the direct involvement of Niamh Lacy-Roberts, the slides were developed by presentations given by Niamh Lacy-Roberts, nlac@food.dtu.dk.