

**NIHR**

Health Protection Research Unit in Healthcare  
Associated Infections and Antimicrobial  
Resistance at University of Oxford



UK Health  
Security  
Agency



MODERNISING  
MEDICAL  
MICROBIOLOGY



NUFFIELD  
DEPARTMENT  
of MEDICINE



UNIVERSITY OF  
OXFORD

# How much is enough for genomic surveillance?

Evaluating genomic diversity in *E. coli* & *Klebsiella* bloodstream  
infection isolates in England from the NEKSUS Study



Monday 19<sup>th</sup> January 2026

Dorottya Nagy, Modernising Medical Microbiology Unit,  
University of Oxford, UK

National *E. coli* and *Klebsiella* bloodstream infection  
and CPE UK Surveillance Study

# Background- BSI

- High **burden** of BSIs in the UK
  - ~**18,000** resistant BSIs 2024/25
  - Increasing since 2020/21
- High **mortality** of AMR BSIs
- **Regional** disparities

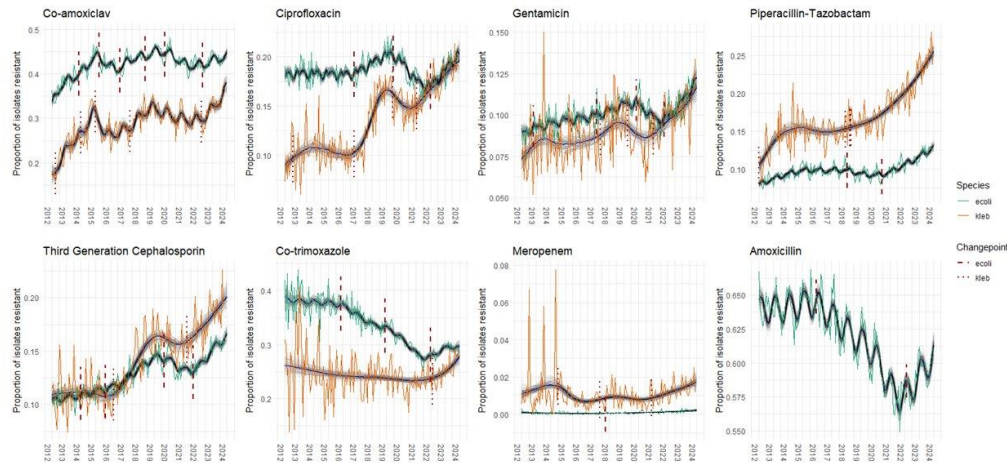
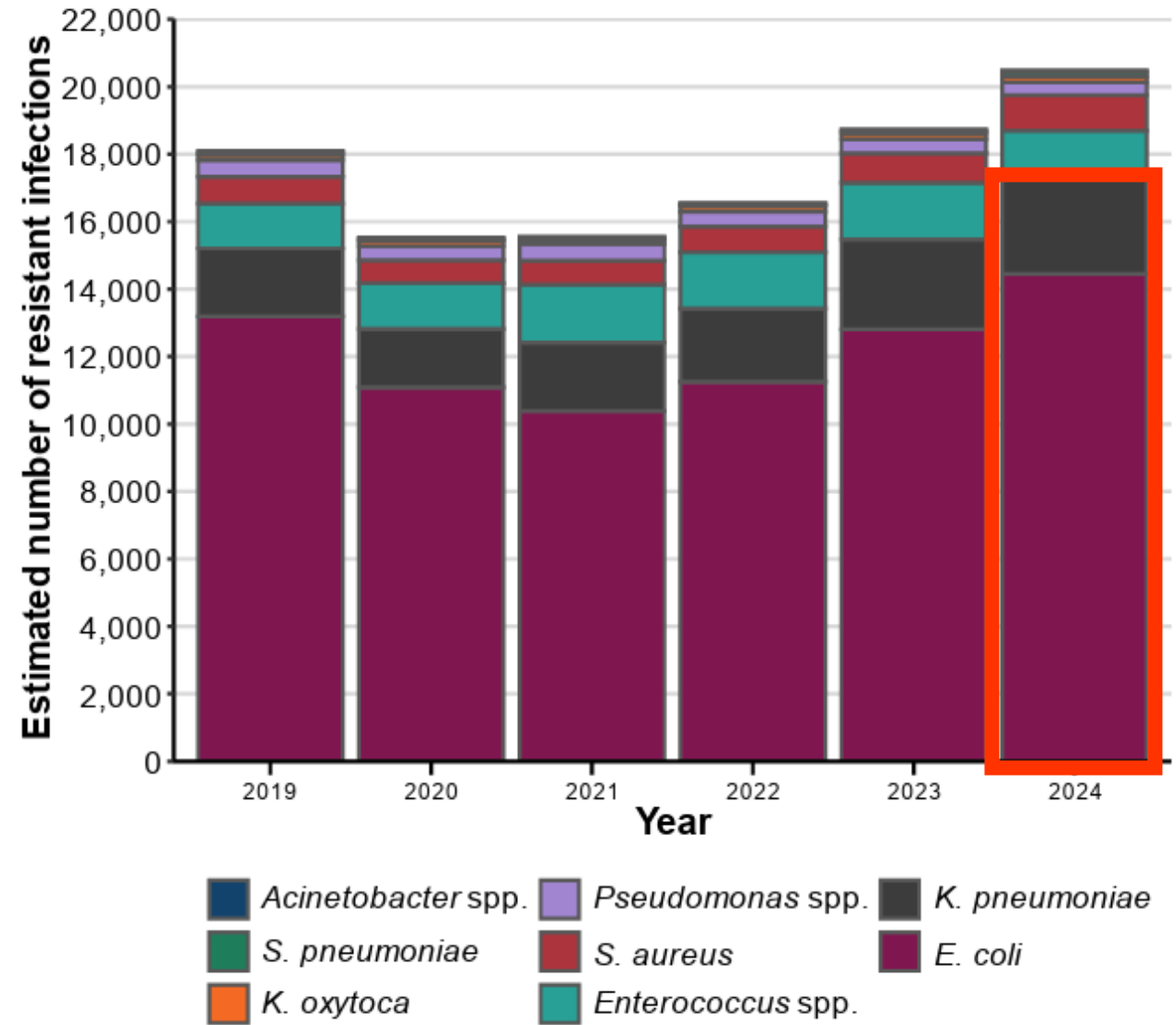


Figure 2.2. Annual estimated total of the burden of antibiotic-resistant bacteraemia episodes, England 2019 to 2024



ESPAUR, 2025

# Methods – The NEKSUS Study

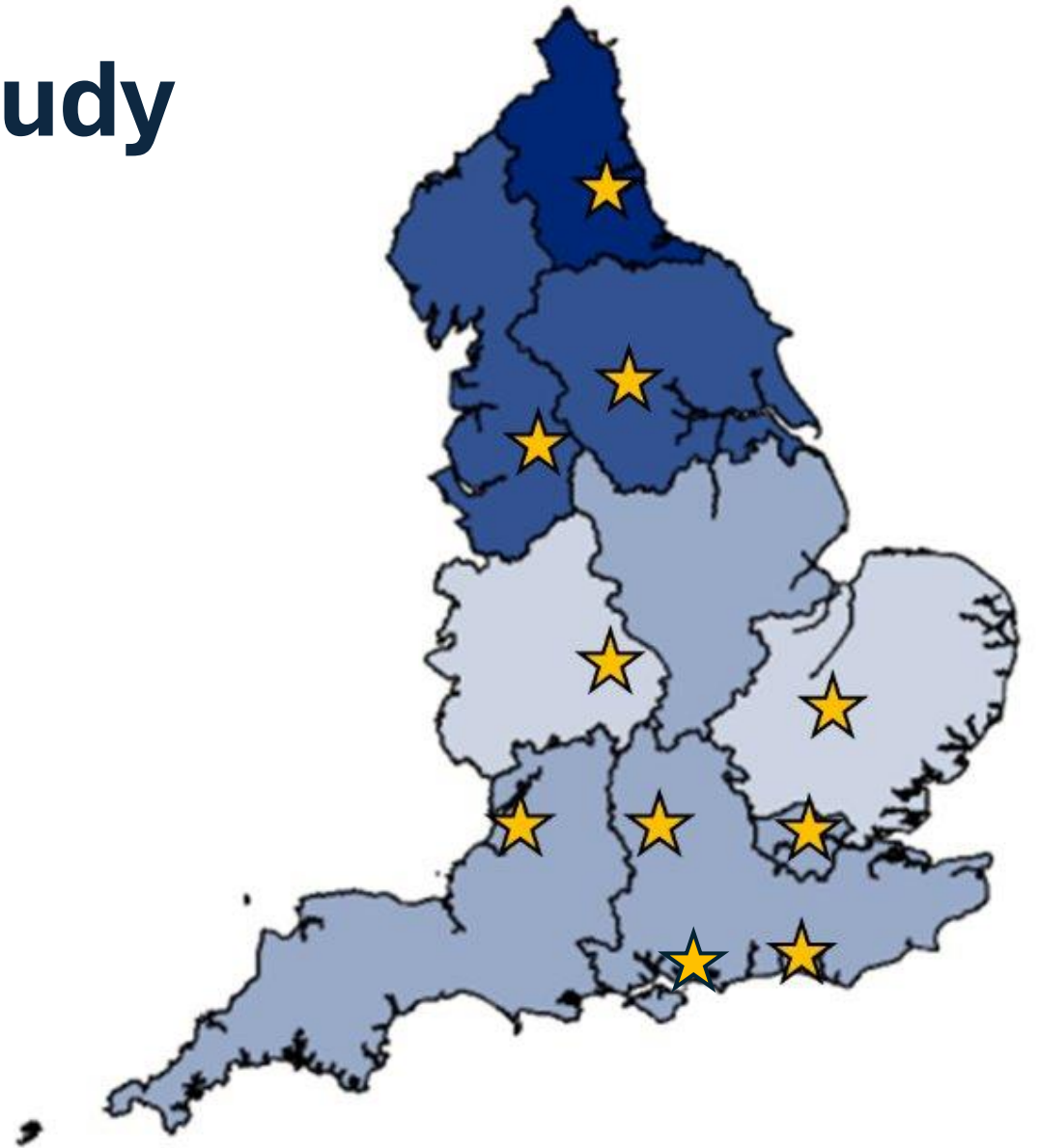
- **10 sites**
  - High turnover acute Trusts
- **Oct 2023 – March 2024:**
  - All *E. coli* and *Klebsiella* spp. BSIs (S + R)
  - CPE positive rectal screens
- **Nanopore sequenced:**
  - **1471** *E. coli*
  - **468** *Klebsiella*
  - ~65% of SGSS submissions at recruited sites

**bioRxiv**  
THE PREPRINT SERVER FOR BIOLOGY

**Nanopore long-read only genome assembly of clinical Enterobacterales isolates is complete and accurate**

 Dorottya Nagy,  Valentina Pennetta,  Gillian Rodger,  Katie Hopkins,  Christopher R. Jones, The NEKSUS Consortium,  Susan Hopkins,  Derrick Crook,  A. Sarah Walker,  Julie Robotham,  Katie L. Hopkins,  Alice Ledda,  David Williams,  Russell Hope,  Colin S. Brown,  Nicole Stoesser,  Samuel Lipworth

doi: <https://doi.org/10.1101/2025.09.15.676237>



# Research Question

How many isolates have to be sequenced to capture a representative proportion of the relevant genetic diversity?

- Species richness
- Common species
- In between?

- MLSTs
- AMR genes (ARGs)
- Plasmids?

# Ecology 101- Measures of diversity

- Species richness

How many unique species are observed in the sample?

- Shannon diversity

$$H' = - \sum_{i=1}^R p_i \ln(p_i)$$

- Simpson diversity

$$D = 1 - \left( \frac{\sum n(n-1)}{N(N-1)} \right)$$

- Sampling coverage

Good's Coverage Estimator

What proportion of the entire community belong to a species that has already been observed in the sample?

$$C = 1 - \frac{f_1}{N}$$

More sensitive to singletons



Less sensitive to singletons

# Ecology 101- Measures of diversity

- Species richness

How many unique species are observed in the sample?

More sensitive to singletons

- Shannon diversity

$$H' = - \sum_{i=1}^R p_i \ln(p_i)$$

- Simpson diversity

$$D = 1 - \left( \frac{\sum n(n-1)}{N(N-1)} \right)$$

- Sampling coverage

Good's Coverage Estimator

What proportion of the entire community belong to a species that has already been observed in the sample?

$$C = 1 - \frac{f_1}{N}$$

Less sensitive to singletons



# Ecology 101- Measures of diversity

- Species richness

How many unique species are observed in the sample?

- Shannon diversity

$$H' = - \sum_{i=1}^R p_i \ln(p_i)$$

- Simpson diversity

$$D = 1 - \left( \frac{\sum n(n-1)}{N(N-1)} \right)$$

- Sampling coverage

Good's Coverage Estimator

What proportion of the entire community belong to a species that has already been observed in the sample?

$$C = 1 - \frac{f_1}{N}$$

More sensitive to singletons



Less sensitive to singletons

# Ecology 101- Measures of diversity

- Species richness

How many unique species are observed in the sample?

- Shannon diversity

$$H' = - \sum_{i=1}^R p_i \ln(p_i)$$

- Simpson diversity

$$D = 1 - \left( \frac{\sum n(n-1)}{N(N-1)} \right)$$

- Sampling coverage

Good's Coverage Estimator

What proportion of the entire community belong to a species that has already been observed in the sample?

$$C = 1 - \frac{f_1}{N}$$

More sensitive to singletons



Less sensitive to singletons

# Ecology 101- Measures of diversity

- Species richness

How many unique species are observed in the sample?

- Shannon diversity

$$H' = - \sum_{i=1}^R p_i \ln(p_i)$$

- Simpson diversity

$$D = 1 - \left( \frac{\sum n(n-1)}{N(N-1)} \right)$$

- Sampling coverage

Good's Coverage Estimator

What proportion of the entire community belong to a species that has already been observed in the sample?

$$C = 1 - \frac{f_1}{N}$$

More sensitive to singletons



Less sensitive to singletons

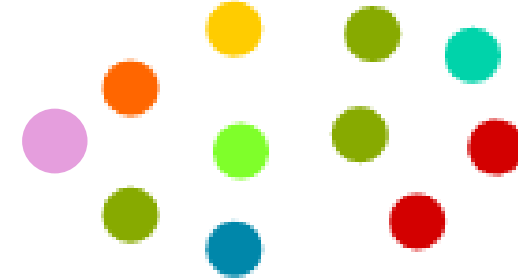
# Ecology 101 - Example

- Species richness

How many unique species are observed in the sample?

7 → 8

+1



- Shannon diversity

1.83 → 1.97

+ 0.14

→ 7 observed species

- Simpson diversity

0.91 → 0.93

+0.02



- Sampling coverage

Good's Coverage Estimator

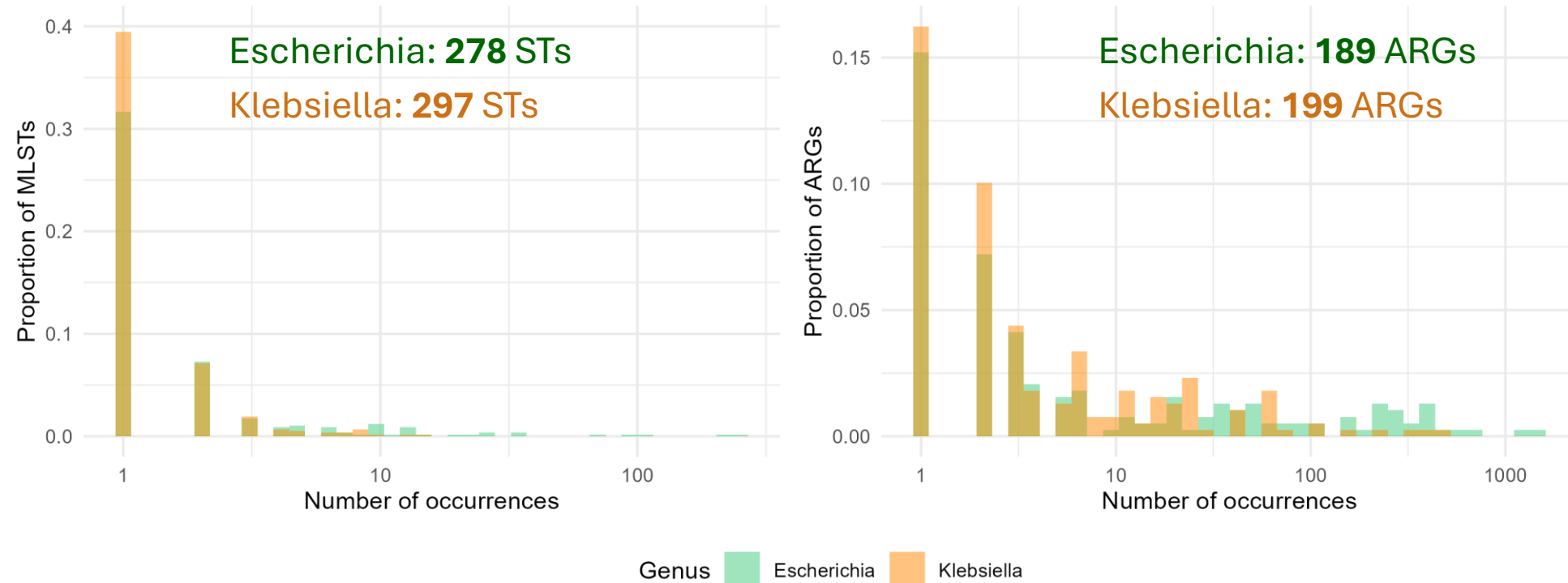
What proportion of the entire community belong to a species that has already been observed in the sample?

0.73 → 0.76

+0.03

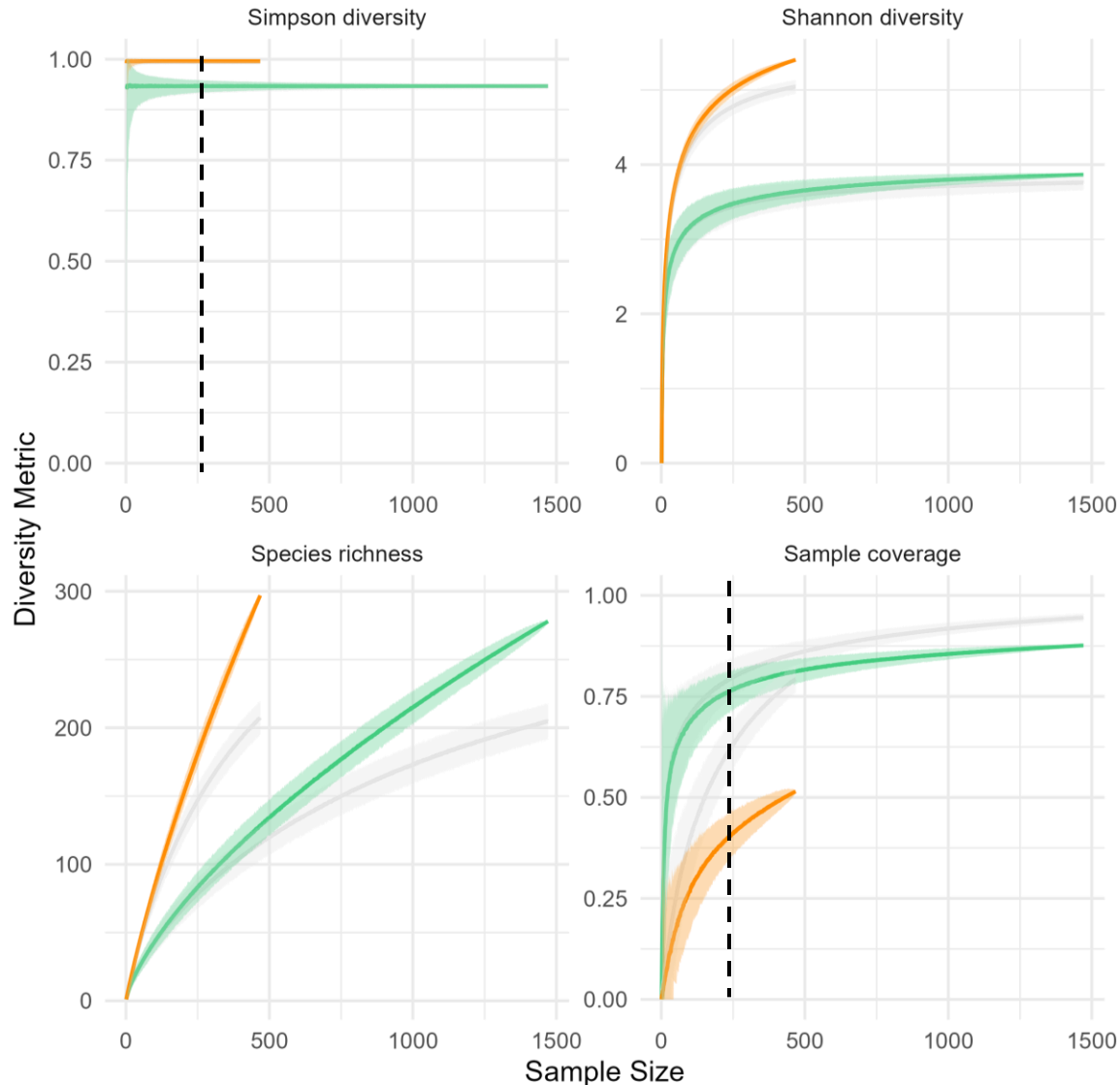


# Results – singletons are common



- Rare MLSTs are more common in *Klebsiella* than *E. coli*
  - MLSTs: **40%** of *Klebsiella* vs **32%** *Escherichia* isolates are singletons
- AMR genes are less skewed
  - **15%** of AMR genes appear in only 1 isolate
- Does this matter?

# Results – Rarefaction of MLST



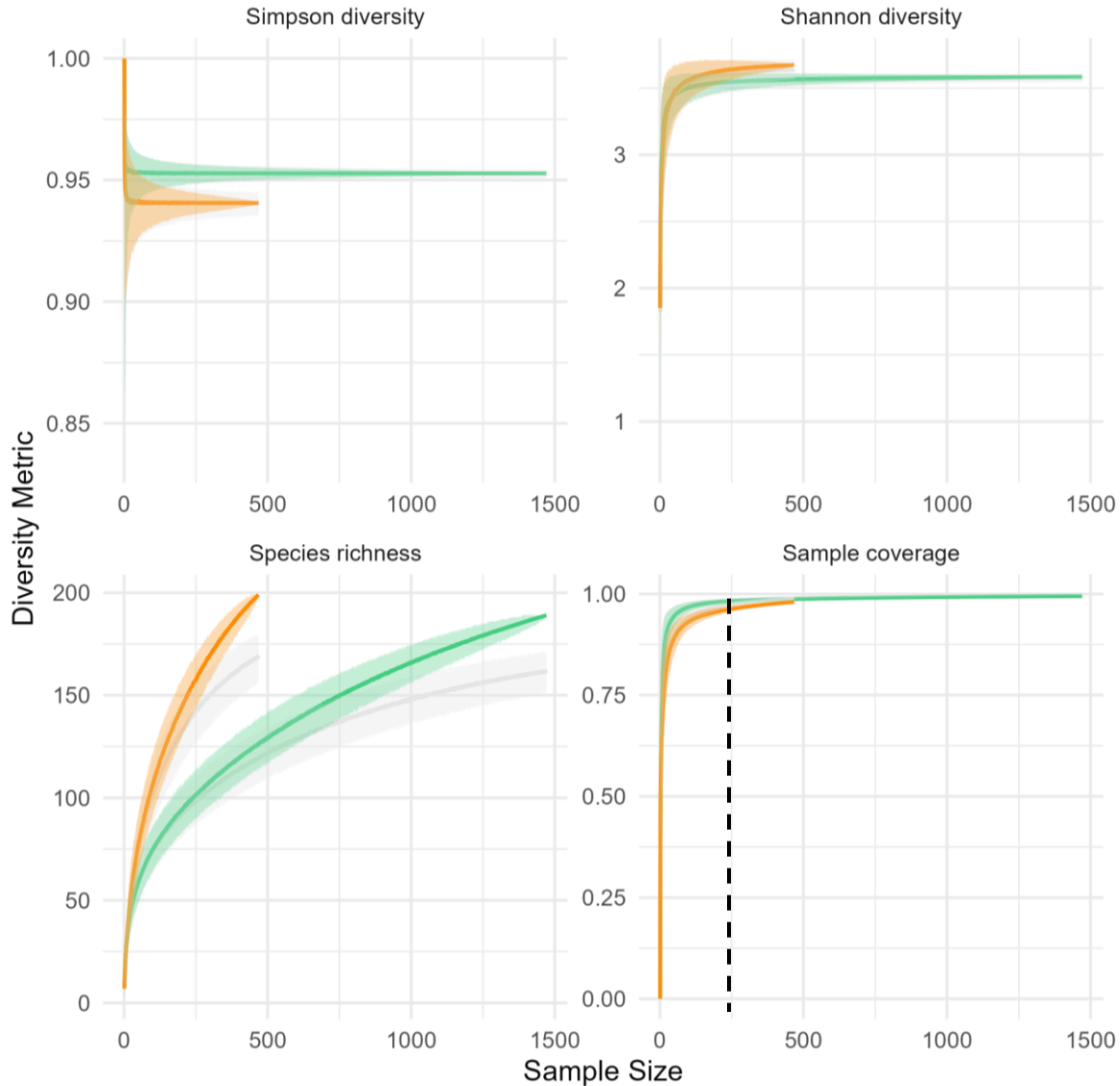
## Measures of diversity

- Species richness is a poor measure
- Simpson diversity is stable
  - Even at small sample sizes (250):
    - Escherichia:  $0.93 \pm 0.017$
    - Klebsiella:  $0.99 \pm 0.0015$
- Sampling without replacement more intuitive

## *Klebsiella* vs *Escherichia*

- More diversity in *Klebsiella*
- Sampling coverage in whole NEKSUS sample:
  - **51%** vs **88%**
- At 250 samples:
  - **41%** (36-46%) vs **77%** (72-82%)

# Results – Rarefaction of ARGs



## Measures of diversity

- Simpson diversity is stable

## *Klebsiella* vs *Escherichia*

- ARG diversity more similar *cf* MLST
- Sampling coverage in whole NEKSUS sample:
  - **98%** vs **99%**
- At 250 samples:
  - **96%** (96-97%) vs **98%** (98-99%)
- Klebsiella has less diversity in common ARGs

# Research Questions 2

What is the minimum sample size,  $n$ , needed to detect the presence vs absence of a lineage that occurs at a **frequency** of  $P_{Vi}$  with a **probability** of  $p$ ?

$$n = \frac{\log(1 - p)}{\log(1 - P_{Vi})}.$$

What is the minimum sample size,  $n$ , needed to estimate prevalence,  $P_{Vi}$ , of a certain lineage with a **precision**,  $d$ , with confidence level,  $Z$  (Z-statistic)?

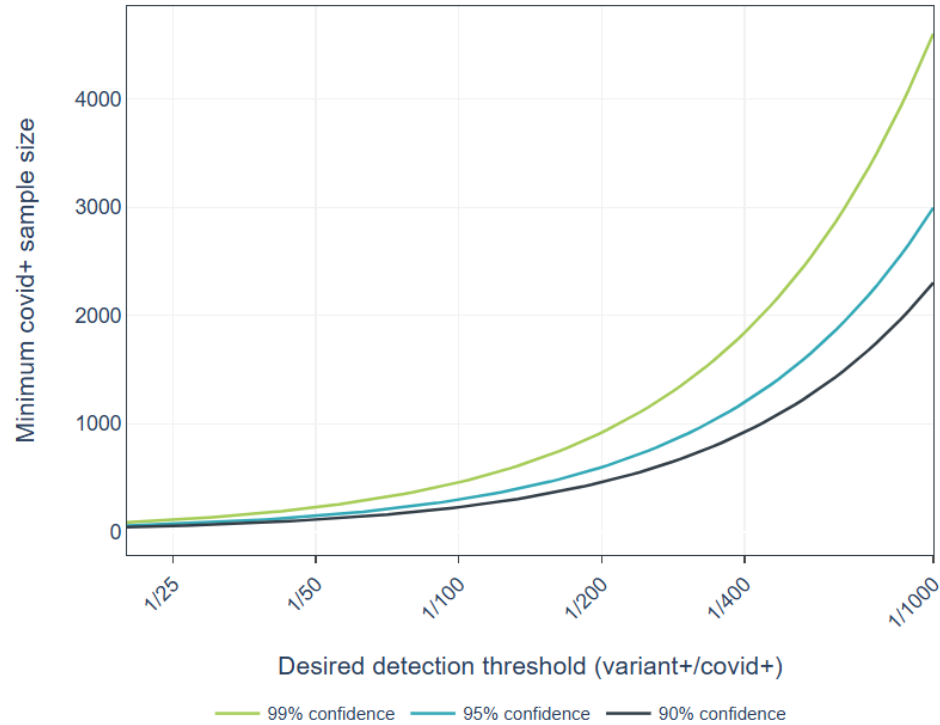
$$n = \frac{Z^2 P_{Vi} (1 - P_{Vi})}{d^2},$$

**Assumptions: The sample is representative and random**

Wohl et al., 2023

# Power calculations

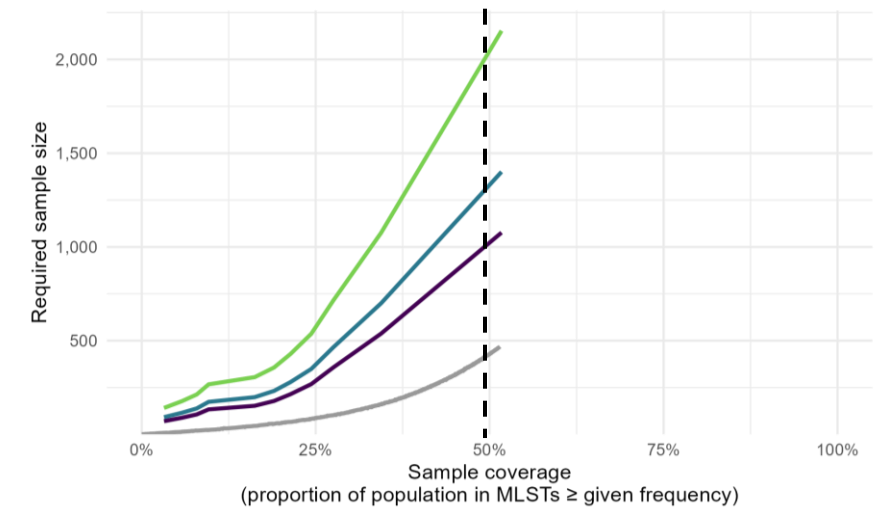
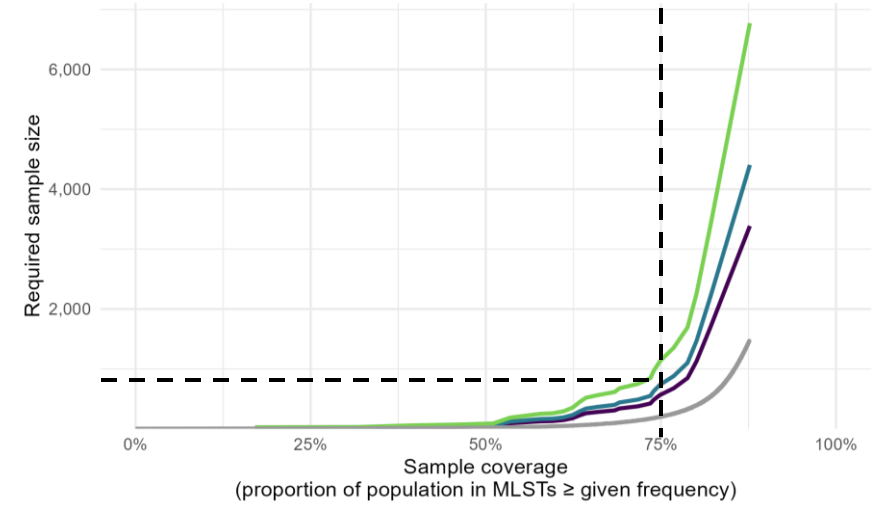
How many SARS-CoV-2 positive specimens should be sequenced to detect a variant?



*E. coli* data



*Klebsiella* data



University of Texas COVID-19 Modelling Consortium sample size calculator: <https://covid-19.tacc.utexas.edu/dashboards/variants/>

# Conclusions

- Many MLST **singletons**
- **Simpson diversity** is a more stable diversity measure than species richness
- **Sampling coverage** may also be useful
- *Klebsiella* has **more MLST** diversity than *E. coli*
- **AMR gene diversity** lower, and more similar between genera
  - **Limited by annotation**

# Future work

- Other genetic features:
  - Clonal clusters (fastBAPS)
  - Plasmids (pling)
- Regional & temporal diversity
- Compare to other sampling estimation methods

# Acknowledgments

## Oxford Team

- Sam Lipworth
- Nicole Stoesser
- Valentina Pennetta
- Gillian Roger
- Katie Hopkins
- Aysha Roohi
- Hieu Thai

## UKHSA Team

- Russel Hope
- Colin Brown
- David Williams
- Chris Jones

## Regional NEKSUS

### Collaborators

- London
- Newcastle
- Leeds
- Manchester
- Birmingham
- Cambridge
- Brighton
- Bristol
- Oxford

