KLEBGEN PROJECT
DATA ANALYSIS REPORT

Moise Christian Jr. Meka

July 18th, 2025

Agenda

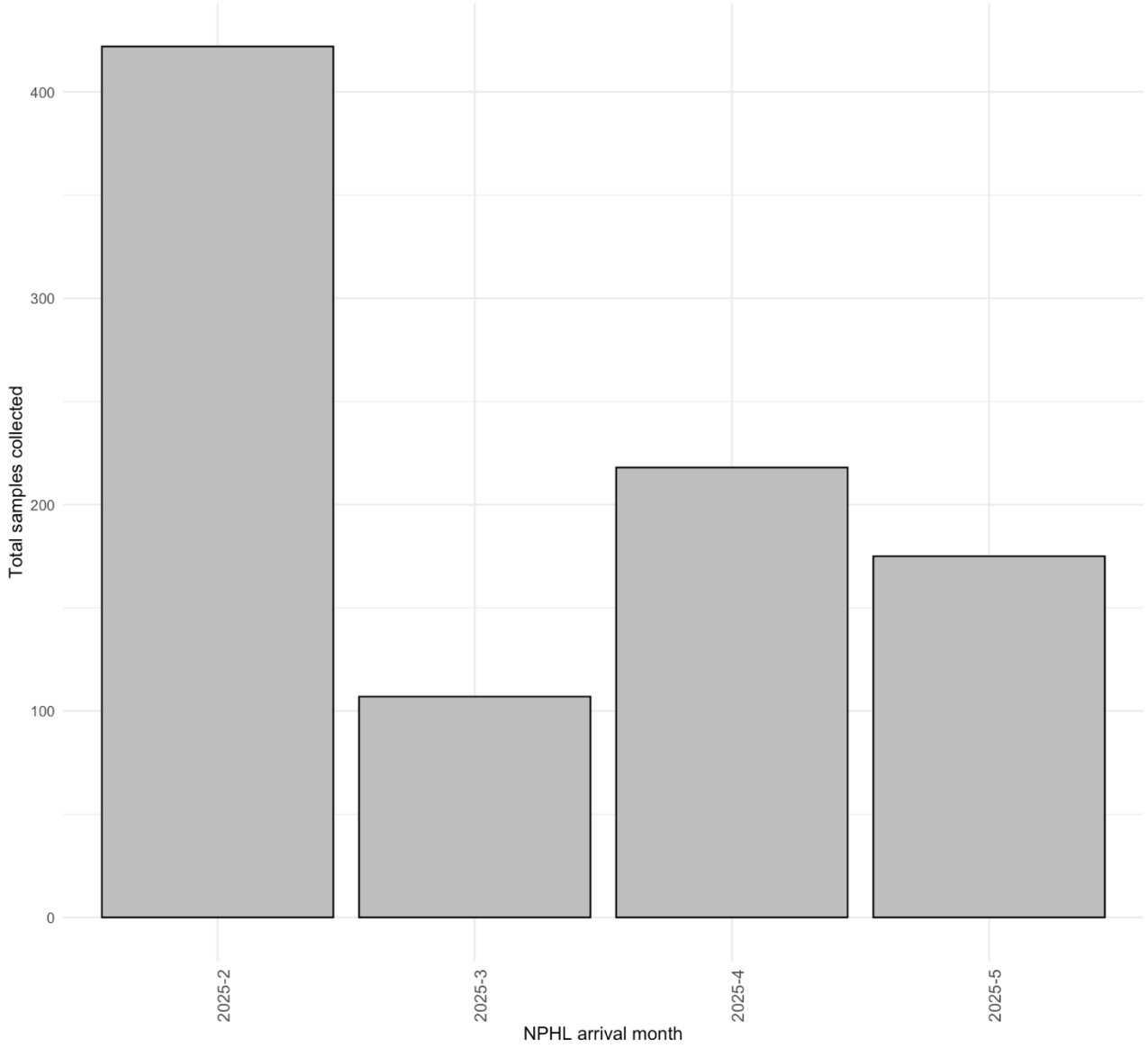
A. Completed activities
B. Samples collection
C. Analysis workflow
D. TheiaProk results
E. *KPSC* analysis results
F. *E. coli/Shigella* analysis results

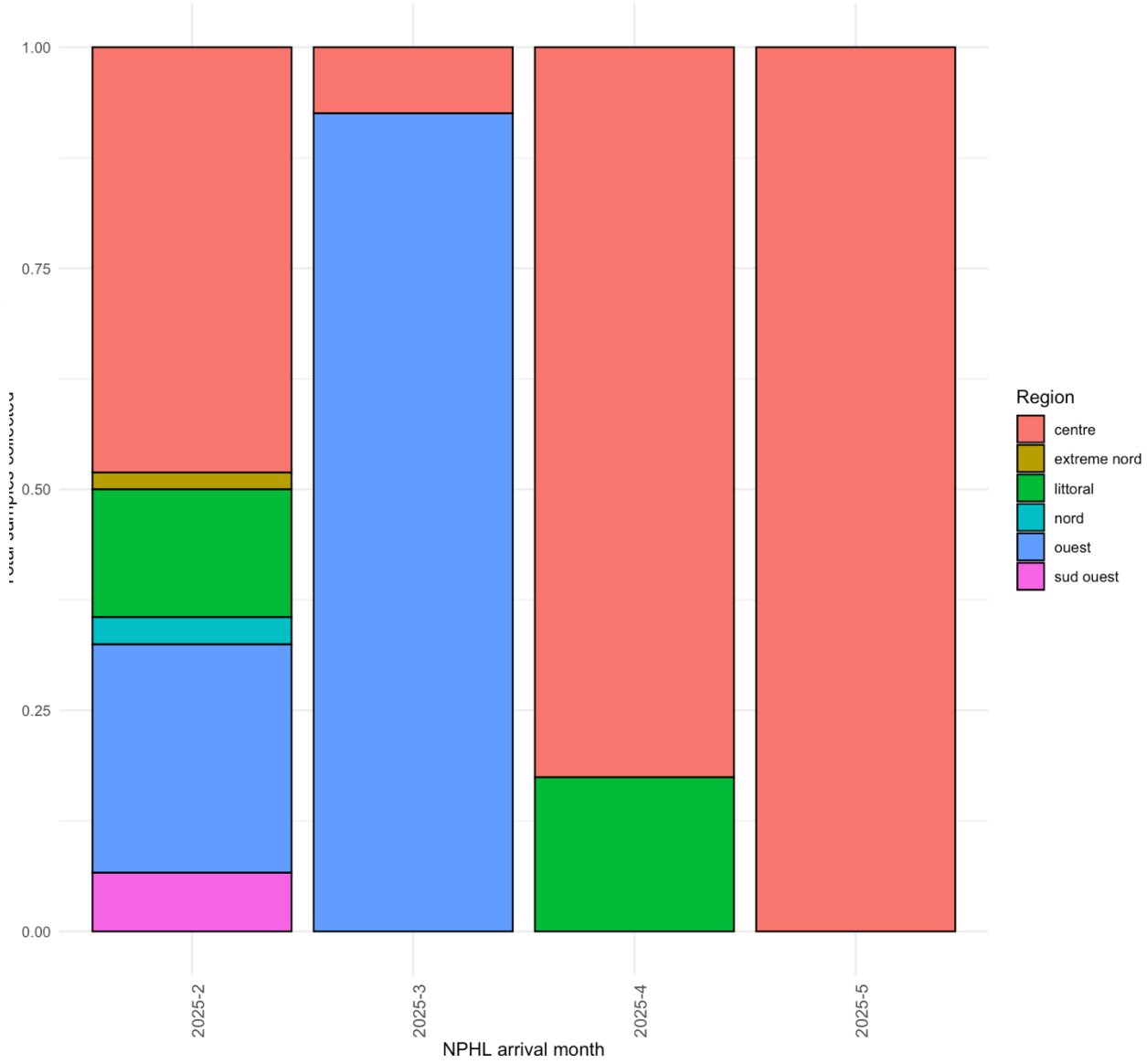
COMPLETED ACTIVITIES

- **Isolates and metadata collection from health facilities**
  - 922 strains : 322 *E. coli*, 498 *Klebsiella* and 102 *Staphylococcus*
  - Sex ratio : 1.09 F/M
  - 82 neonates' strains : 26 *E. coli*, 43 *Klebsiella* and 13 *Staphylococcus*
  - Blood samples : 19.87%
- **Microbiology species verification**
  - 427 strains : 180 *E. coli*, 247 *Klebsiella*
- **Sequencing**
  - 379 sequences : 187 *Klebsiella* and 162 *E. coli*
- **Sharing sequencing results and metadata in Terra**
  - 293 sequences : 127 *E. coli* and 166 *Klebsiella*

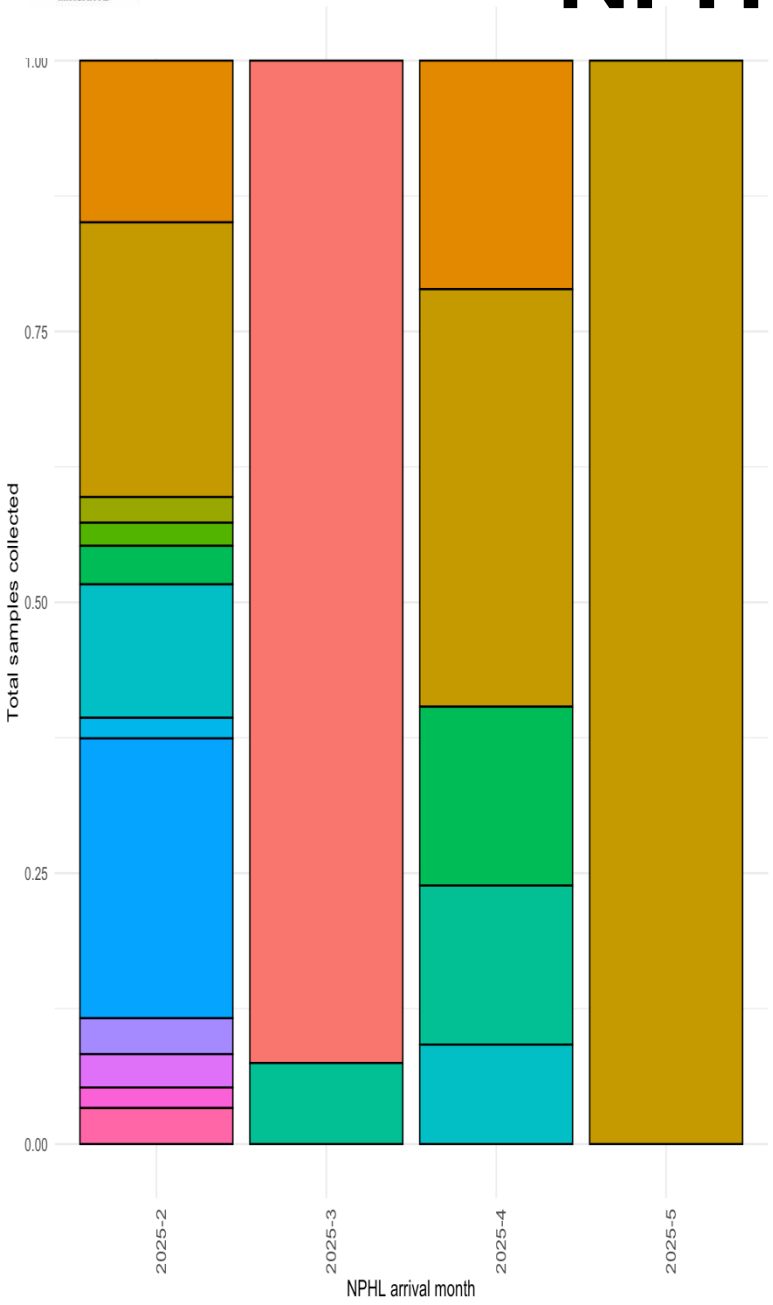
SAMPLES COLLECTION

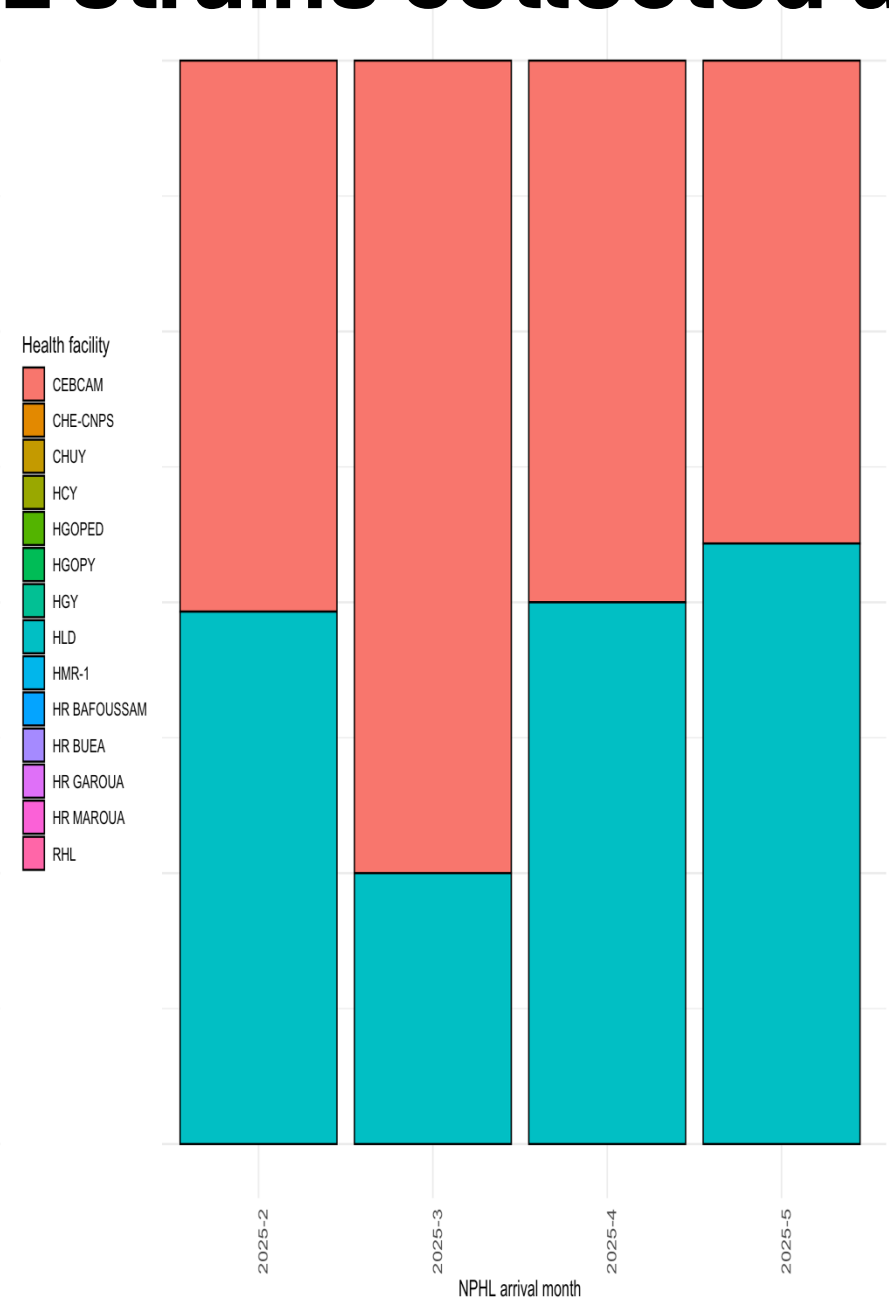
NPHL strains collection dynamics

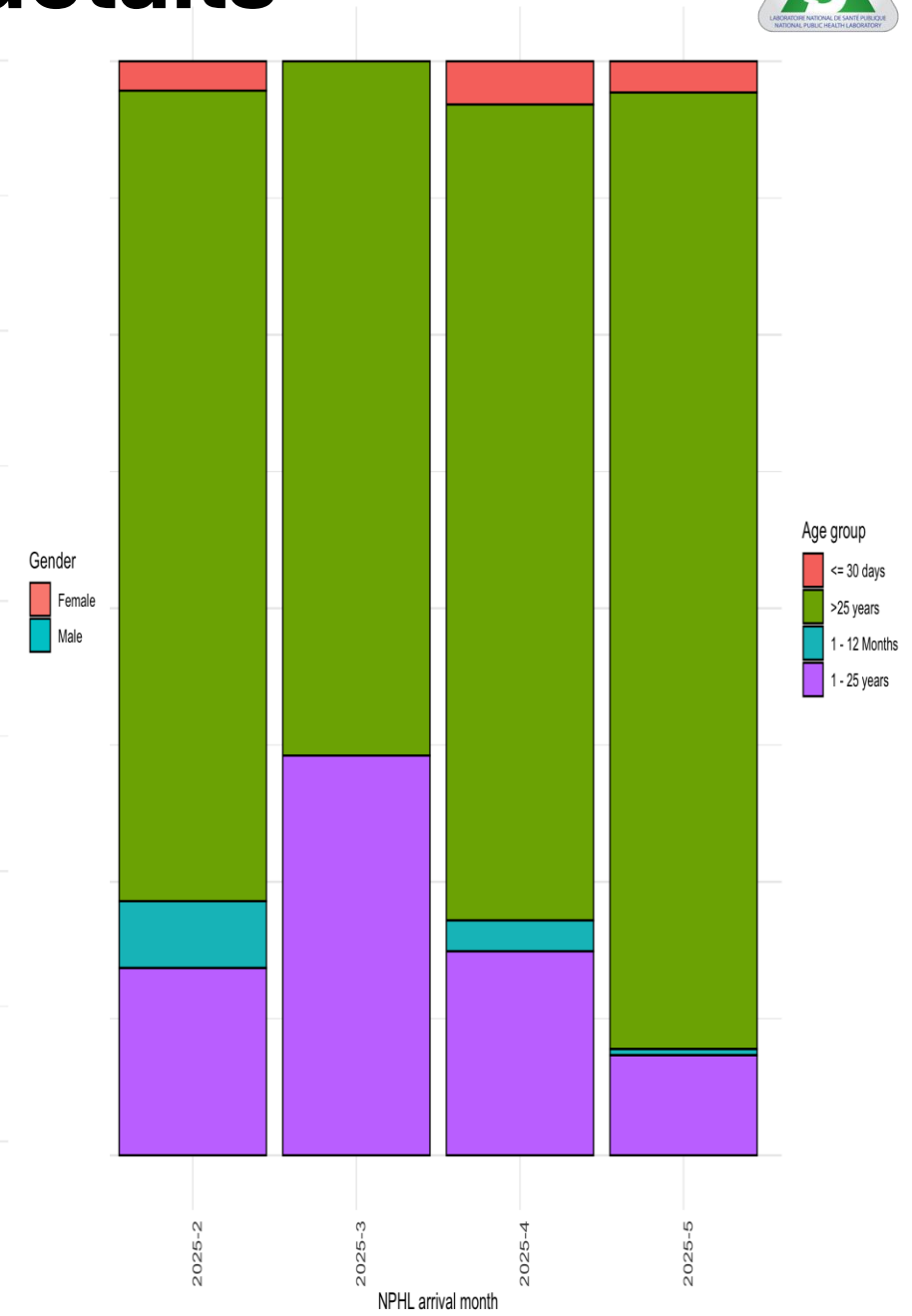




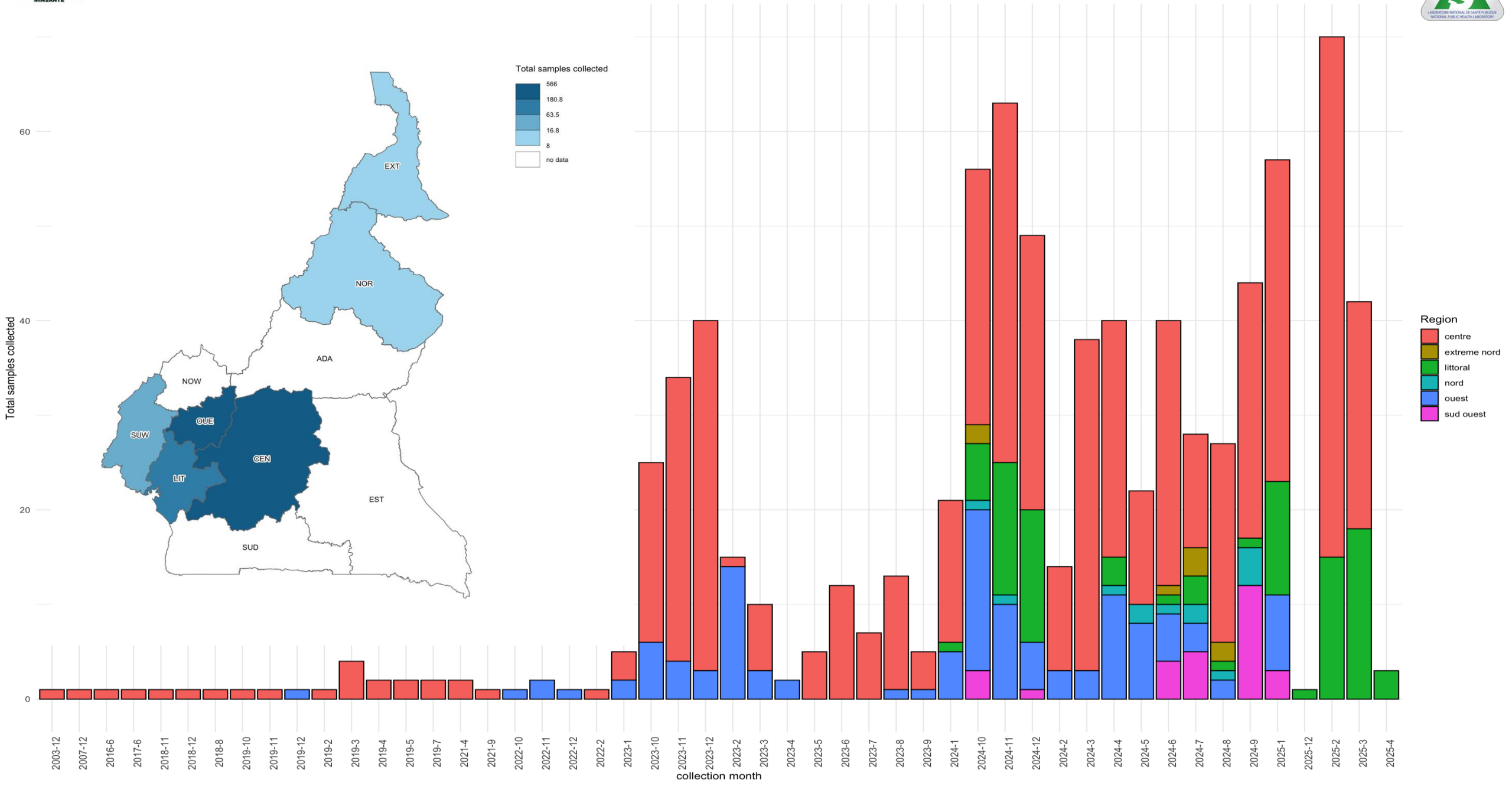
NPHL strains collected details



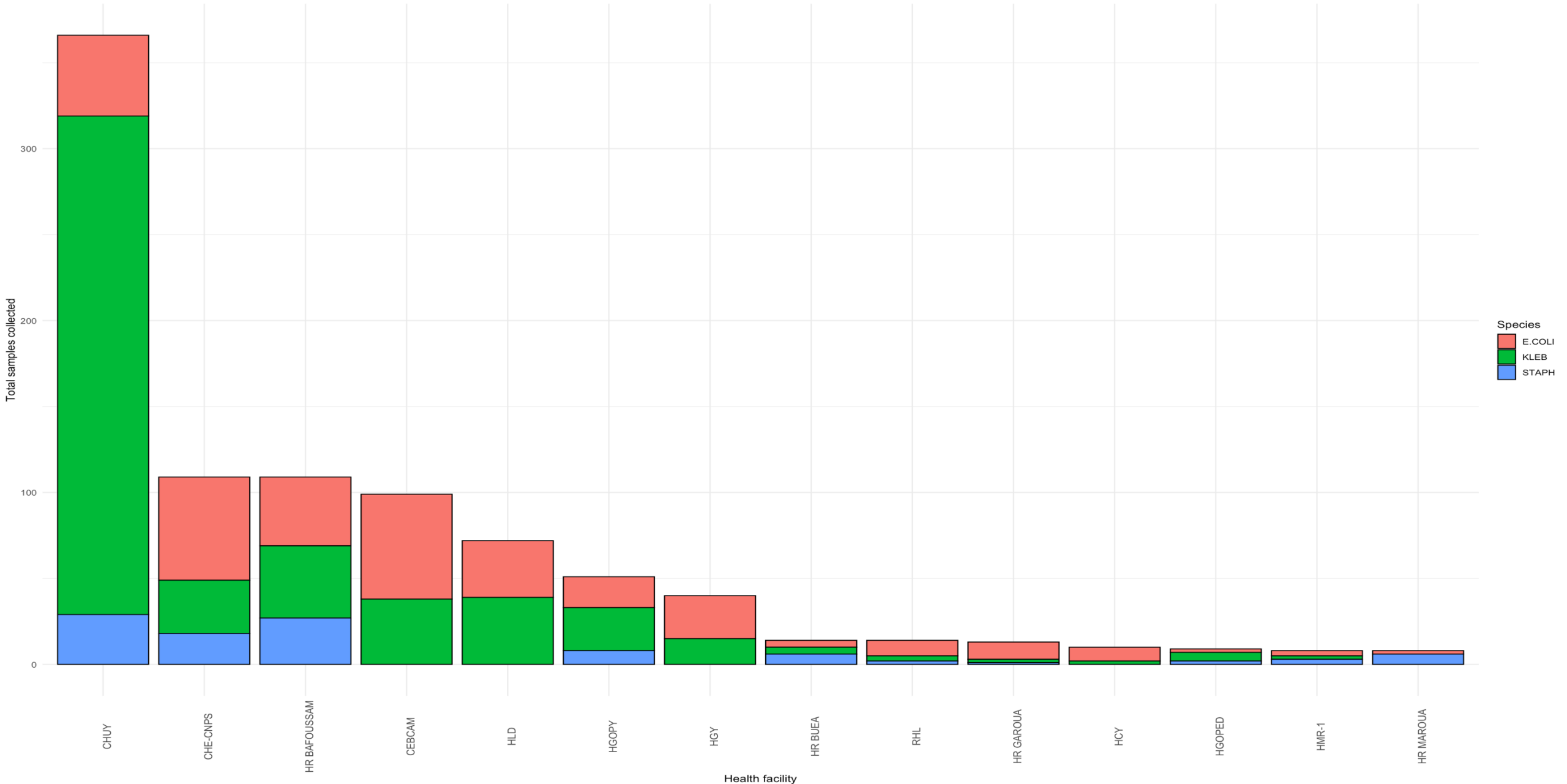




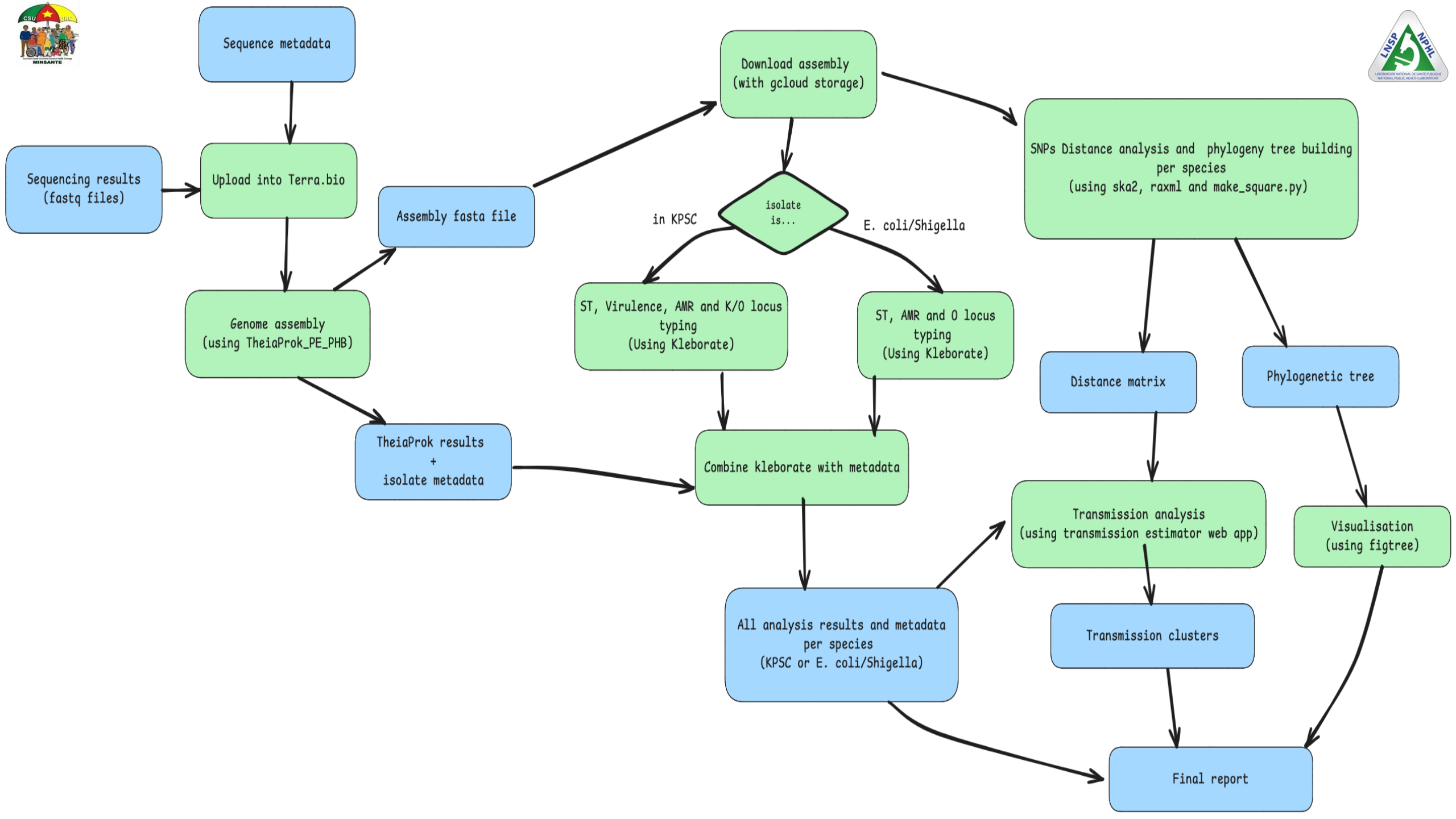
Regional sample collection dynamics



Isolates distribution per Health facility and strain

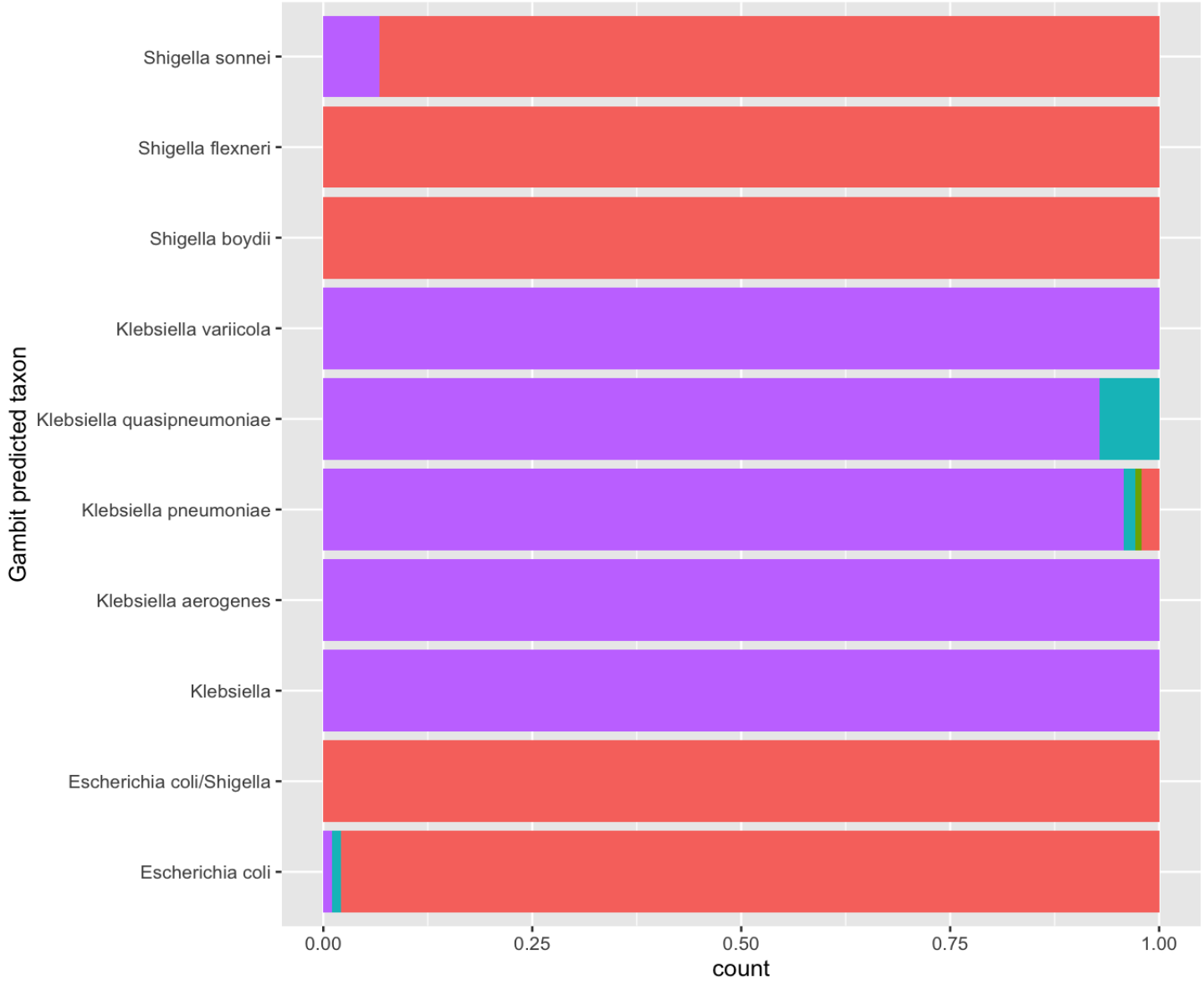


ANALYSIS WORKFLOW


Sequence metadata
Sequencing results (fastq files)
Upload into Terra.bio
Assembly fasta file
Download assembly (with gcloud storage)
SNPs Distance analysis and phylogeny tree building per species (using ska2, raxml and make_square.py)
isolate is...
in KPSC
E. coli/Shigella
Genome assembly (using TheiaProk_PE_PHB)
ST, Virulence, AMR and K/O locus typing (Using Kleborate)
ST, AMR and O locus typing (Using Kleborate)
Distance matrix
Phylogenetic tree
TheiaProk results + isolate metadata
Combine kleborate with metadata
Transmission analysis (using transmission estimator web app)
Visualisation (using figtree)
All analysis results and metadata per species (KPSC or E. coli/Shigella)
Transmission clusters
Final report

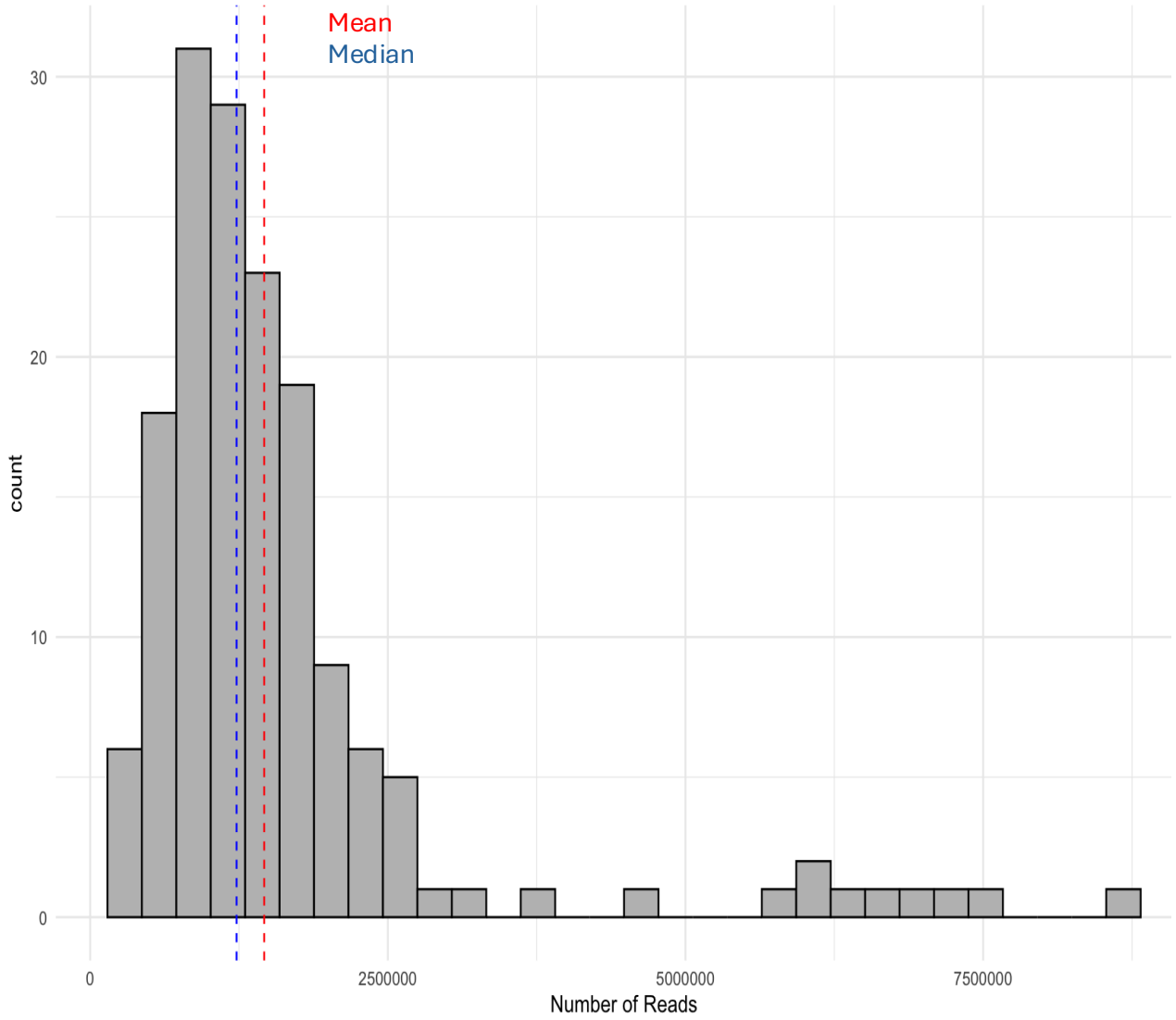
THEIAPROK RESULTS

Comparison of TheiaProk and Microbiology Results

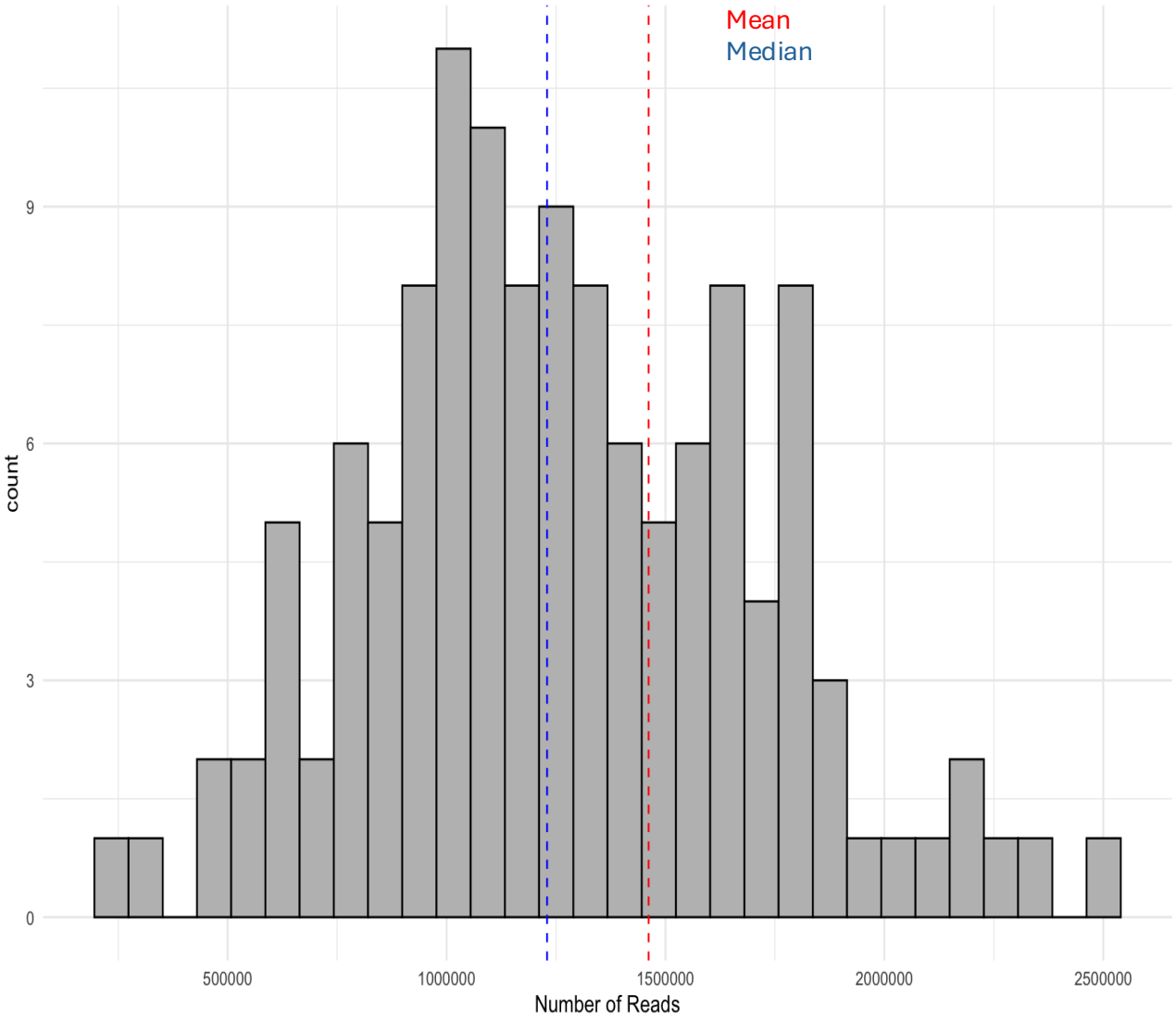


- 287 out of 293 successfully analyzed
  - 6 sequences failed TheiaProk QC
- 93.86% match with microbiology results

Raw sequence quality (Number of reads)

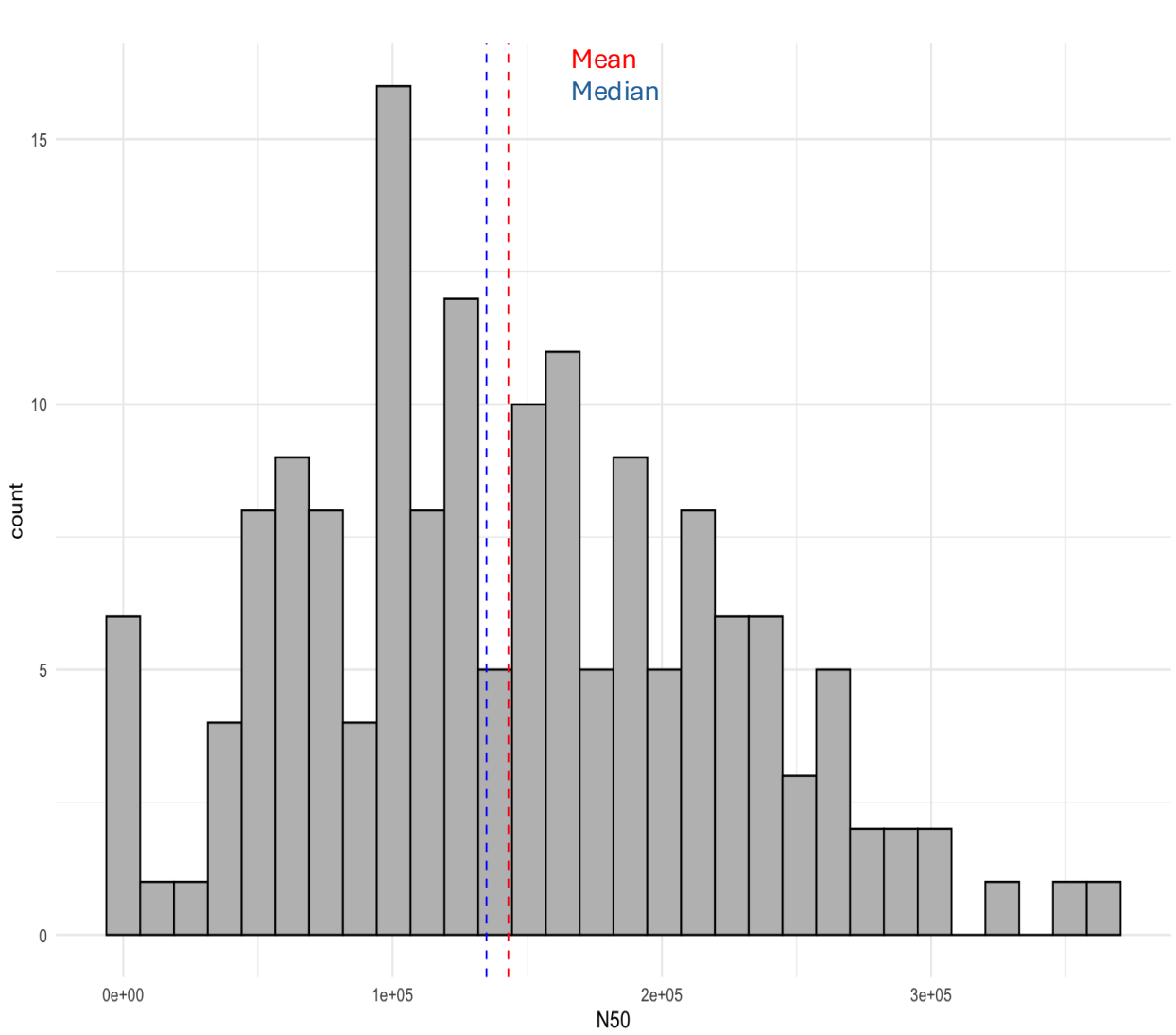


a) *KPSC* Sequences

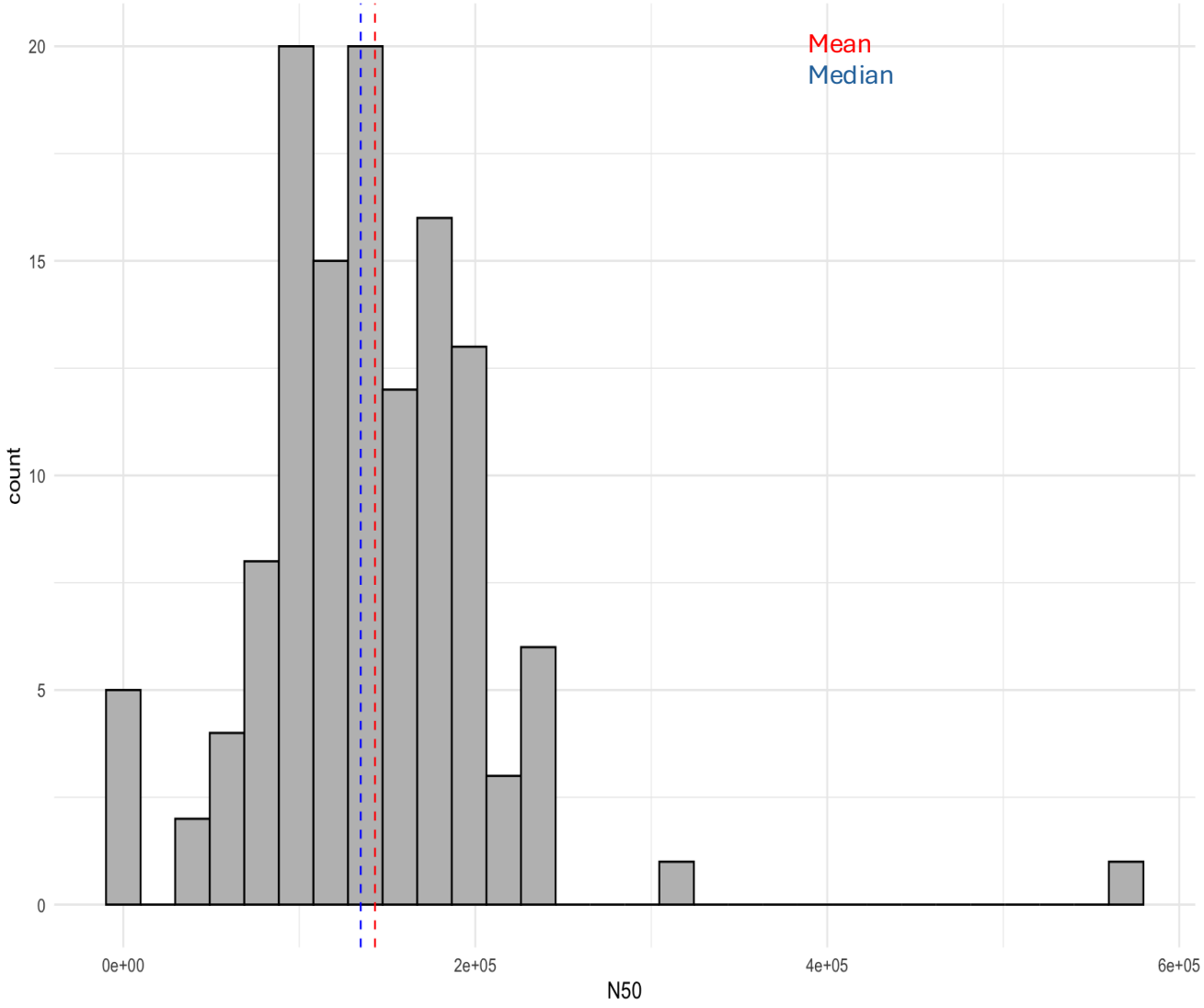


b) *E. coli/Shiguella* Sequences

Sequence assembly quality (N50)



a) *KPSC* Sequences

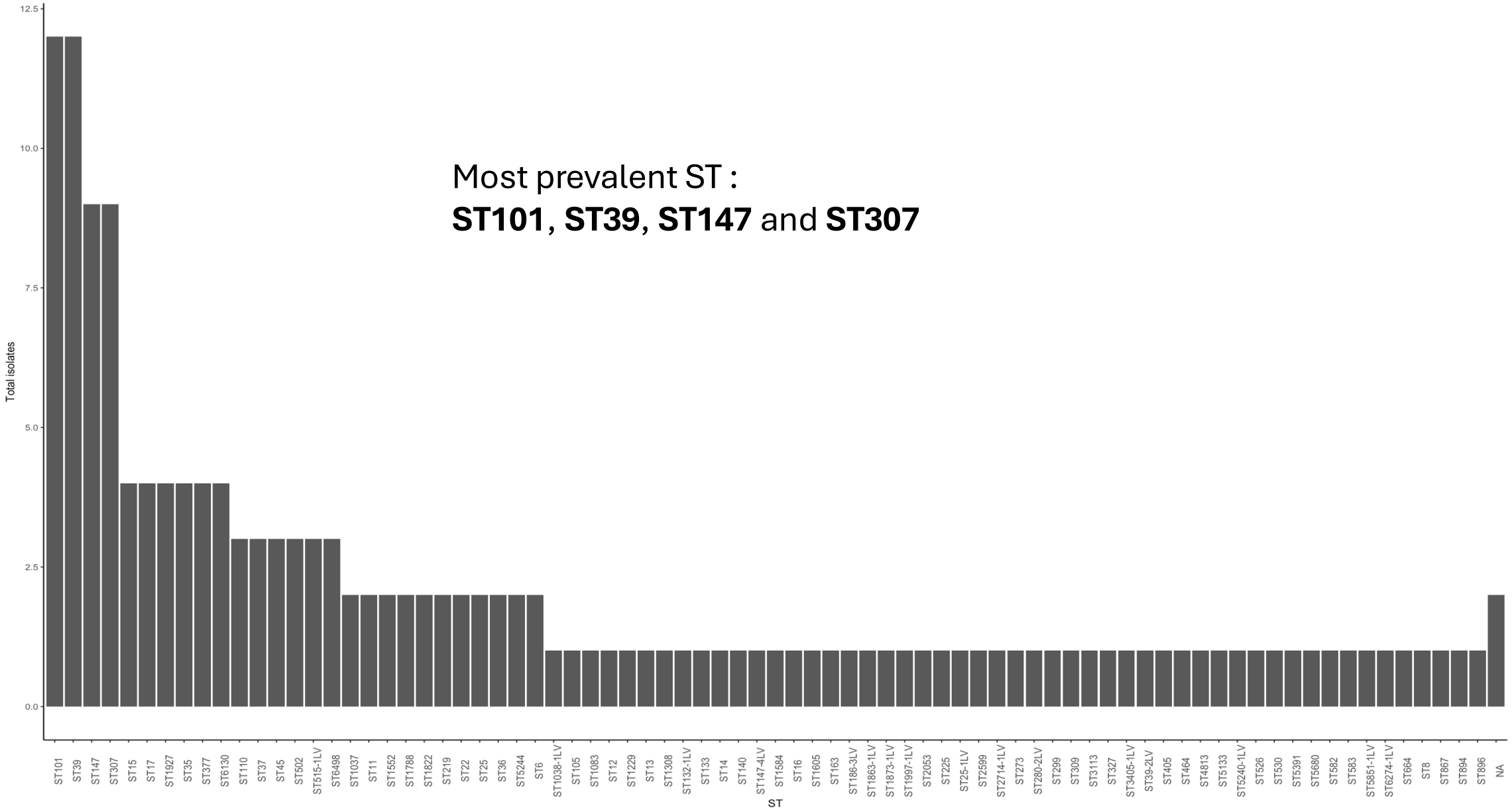


b) *E. coli/Shiguella* Sequences

KPSC ANALYSIS RESULTS

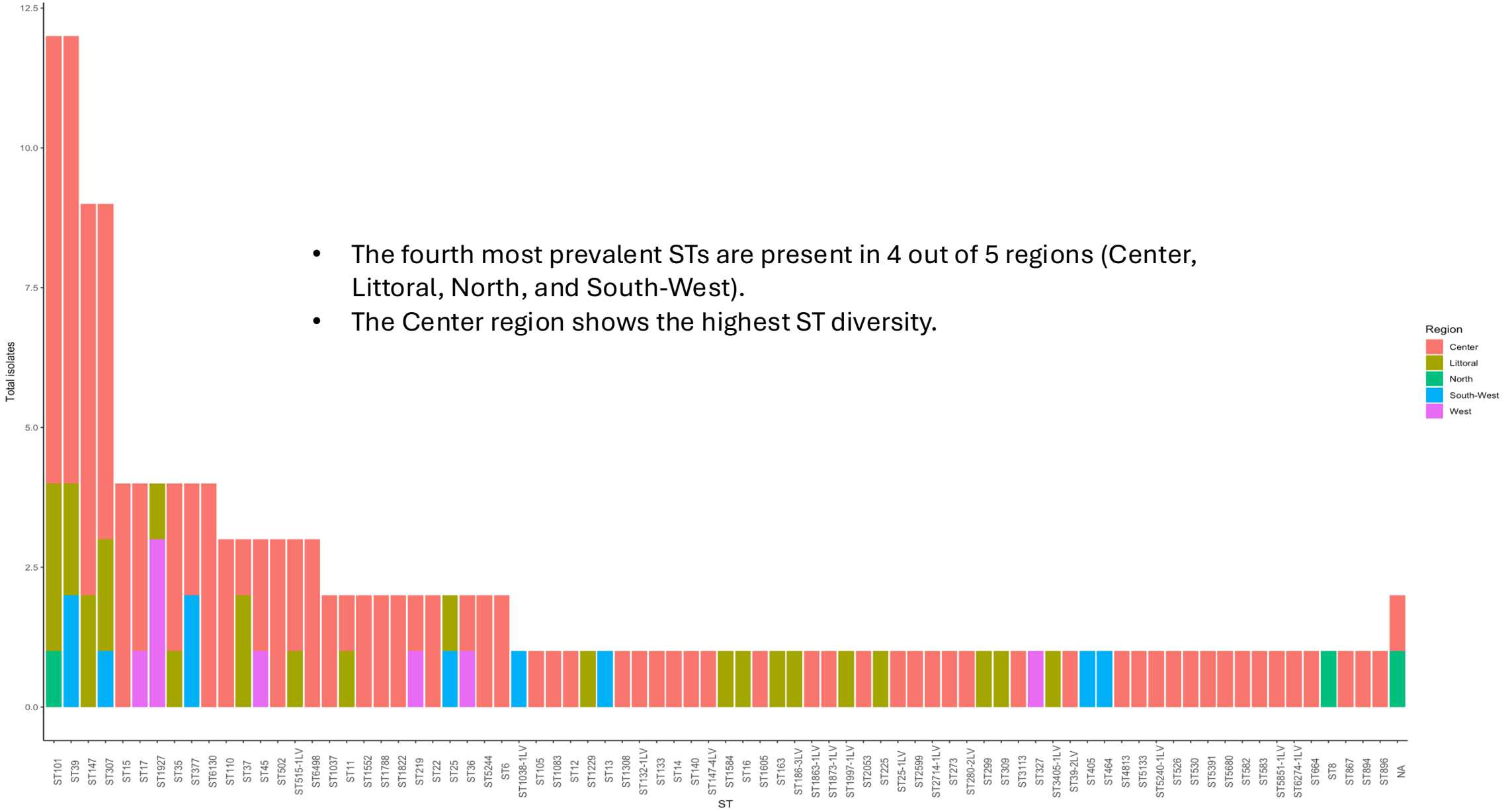
ST prevalence distribution

Most prevalent ST :
ST101, **ST39**, **ST147** and **ST307**

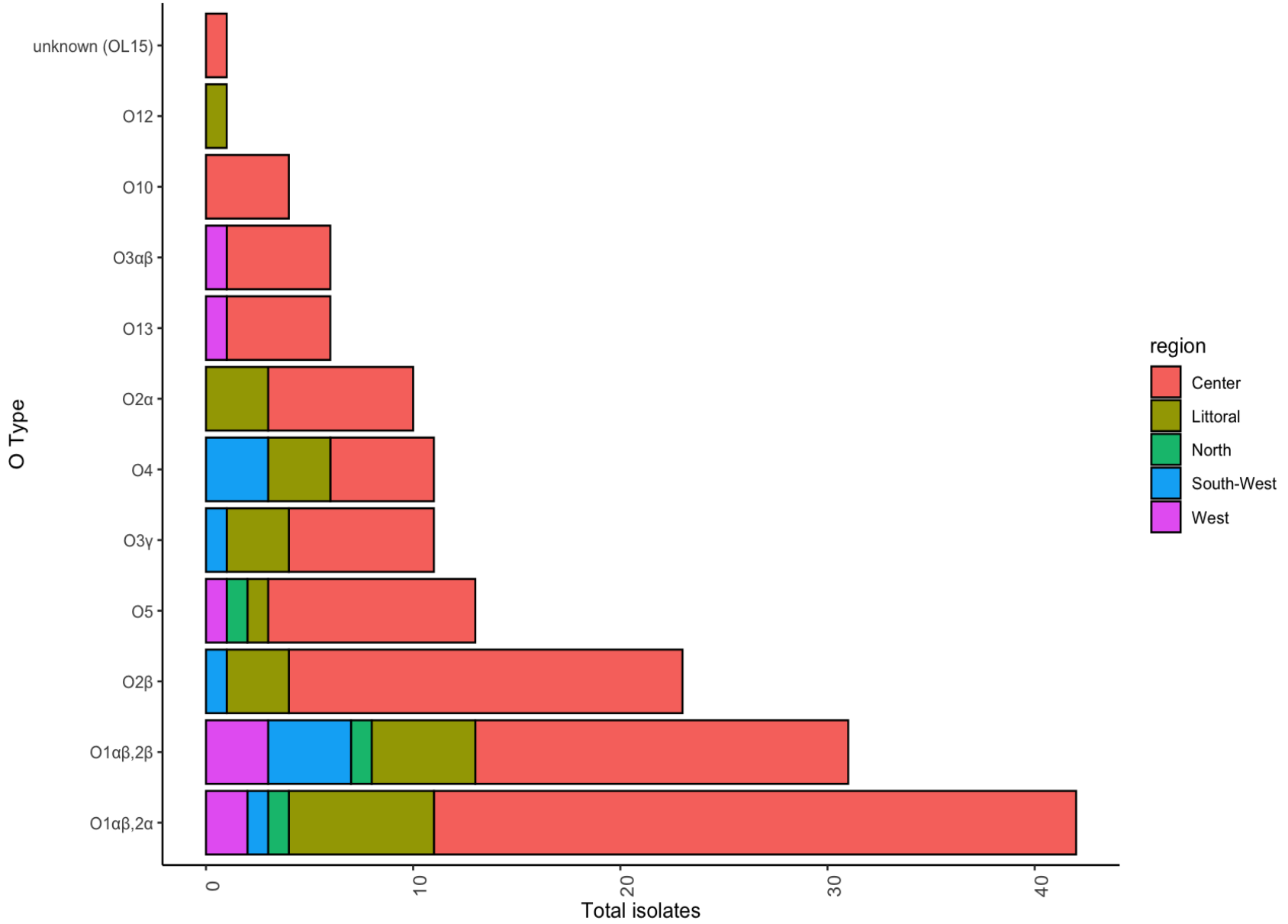


ST prevalence distribution per region

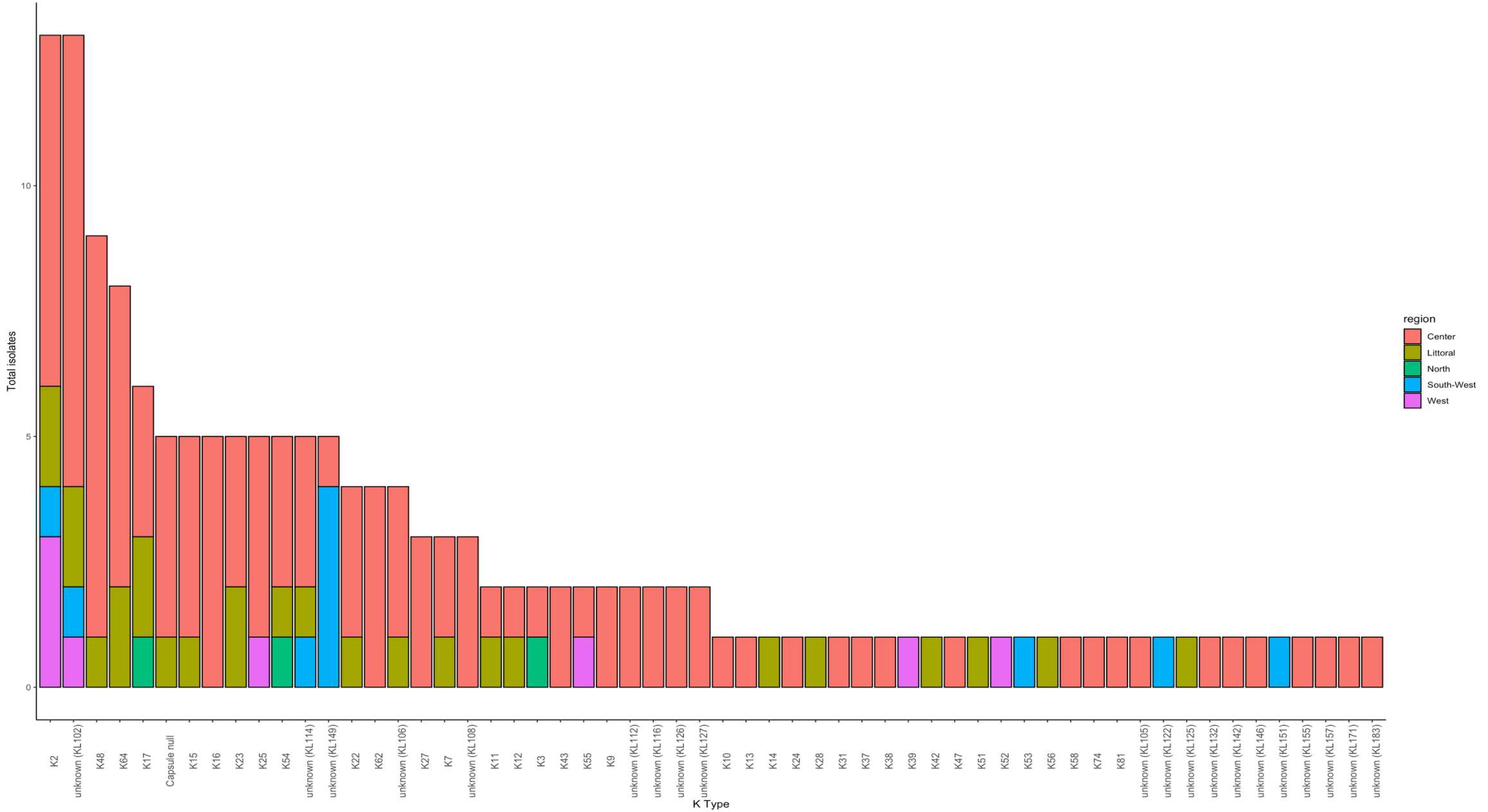
- The fourth most prevalent STs are present in 4 out of 5 regions (Center, Littoral, North, and South-West).
- The Center region shows the highest ST diversity.



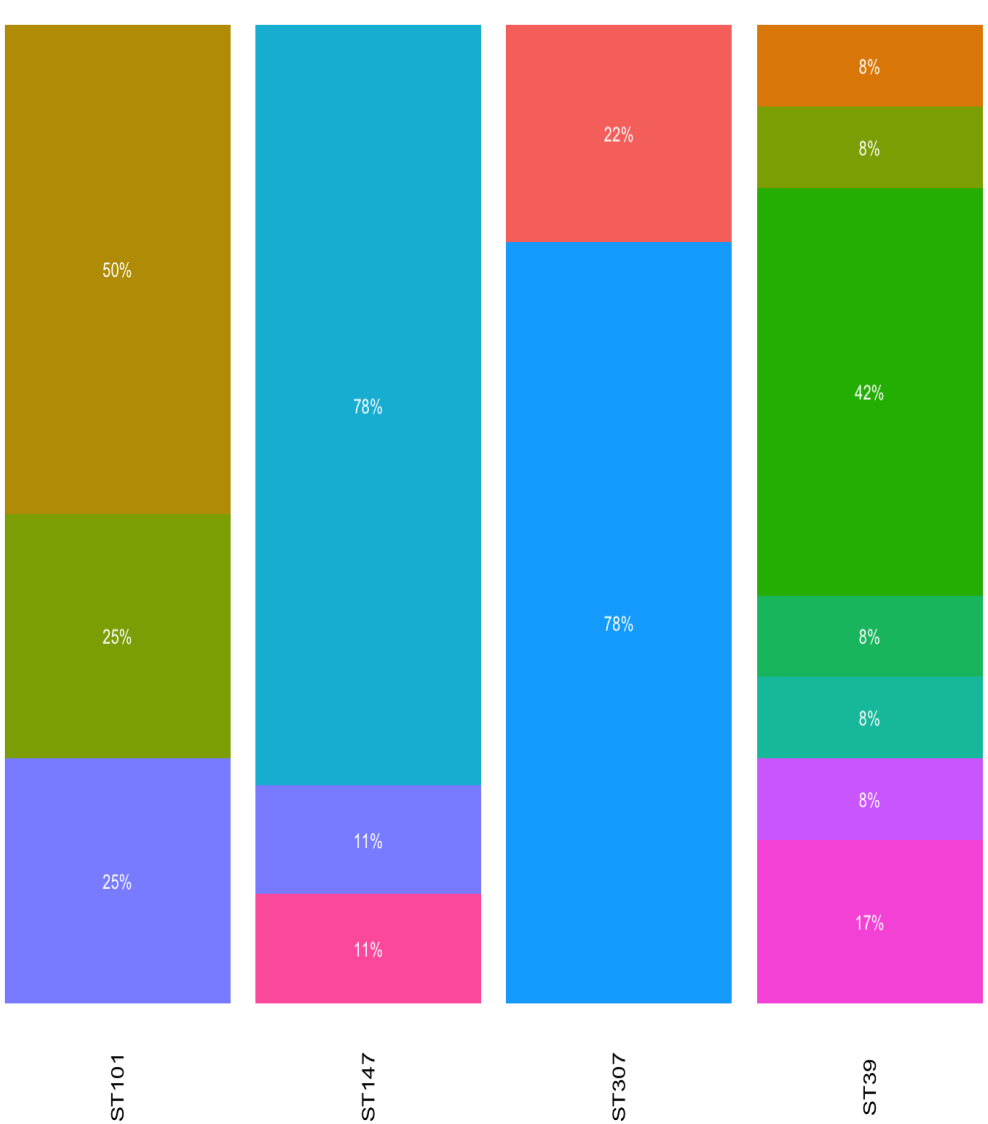
O type Distribution per region



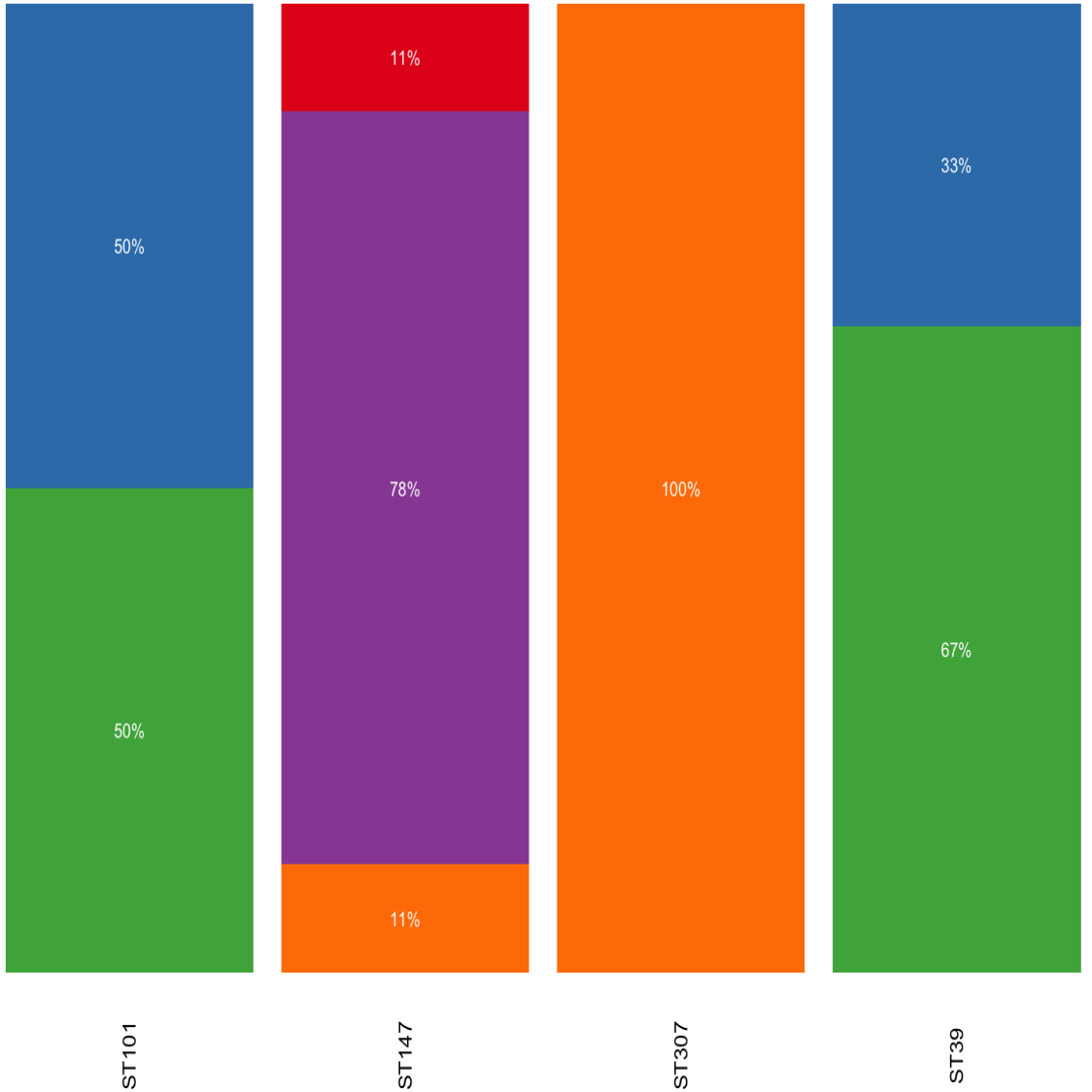
K type Distribution per region



Most prevalent ST : K and O type Distribution



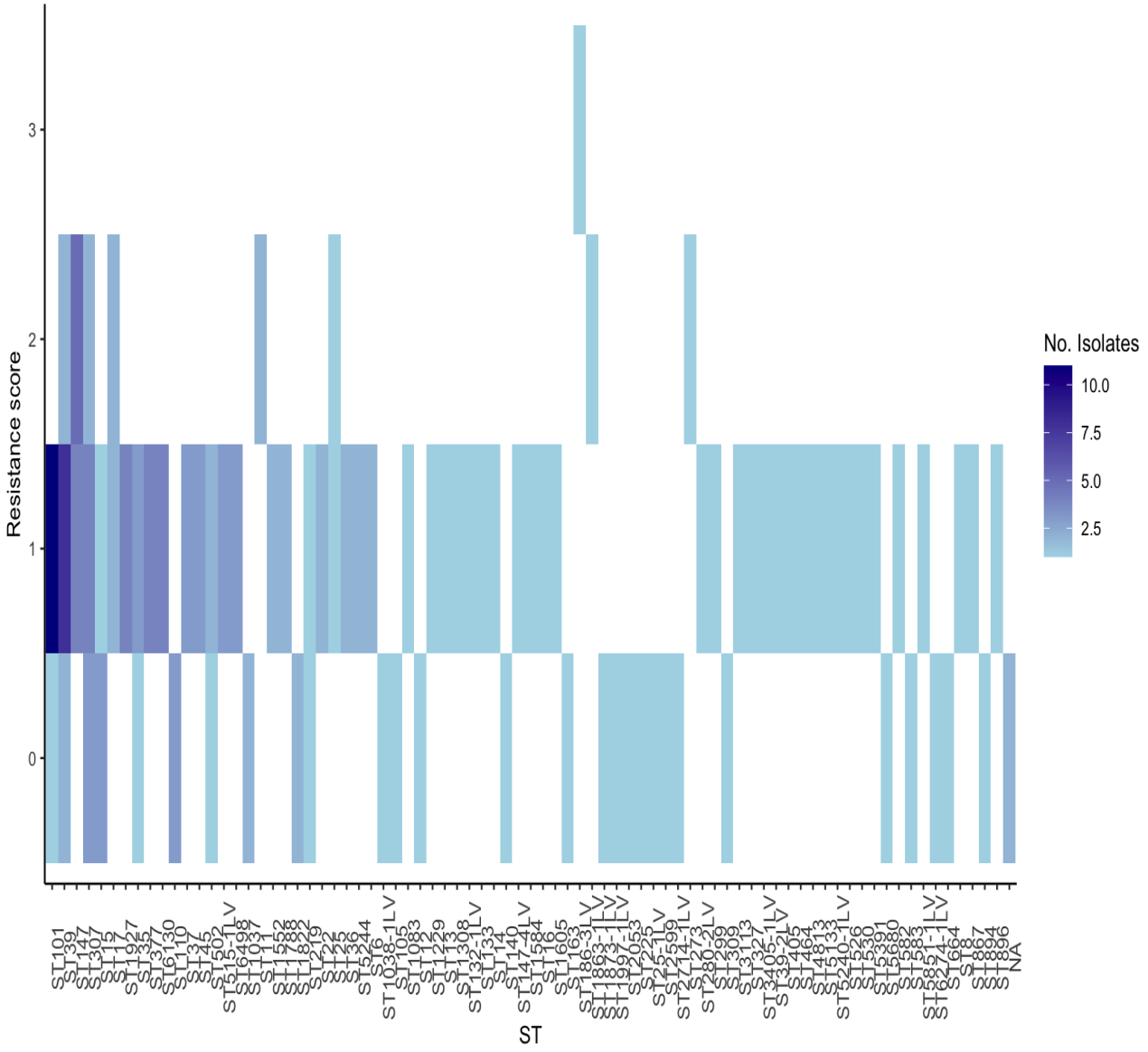
a) Distribution of K type within most prevalent ST



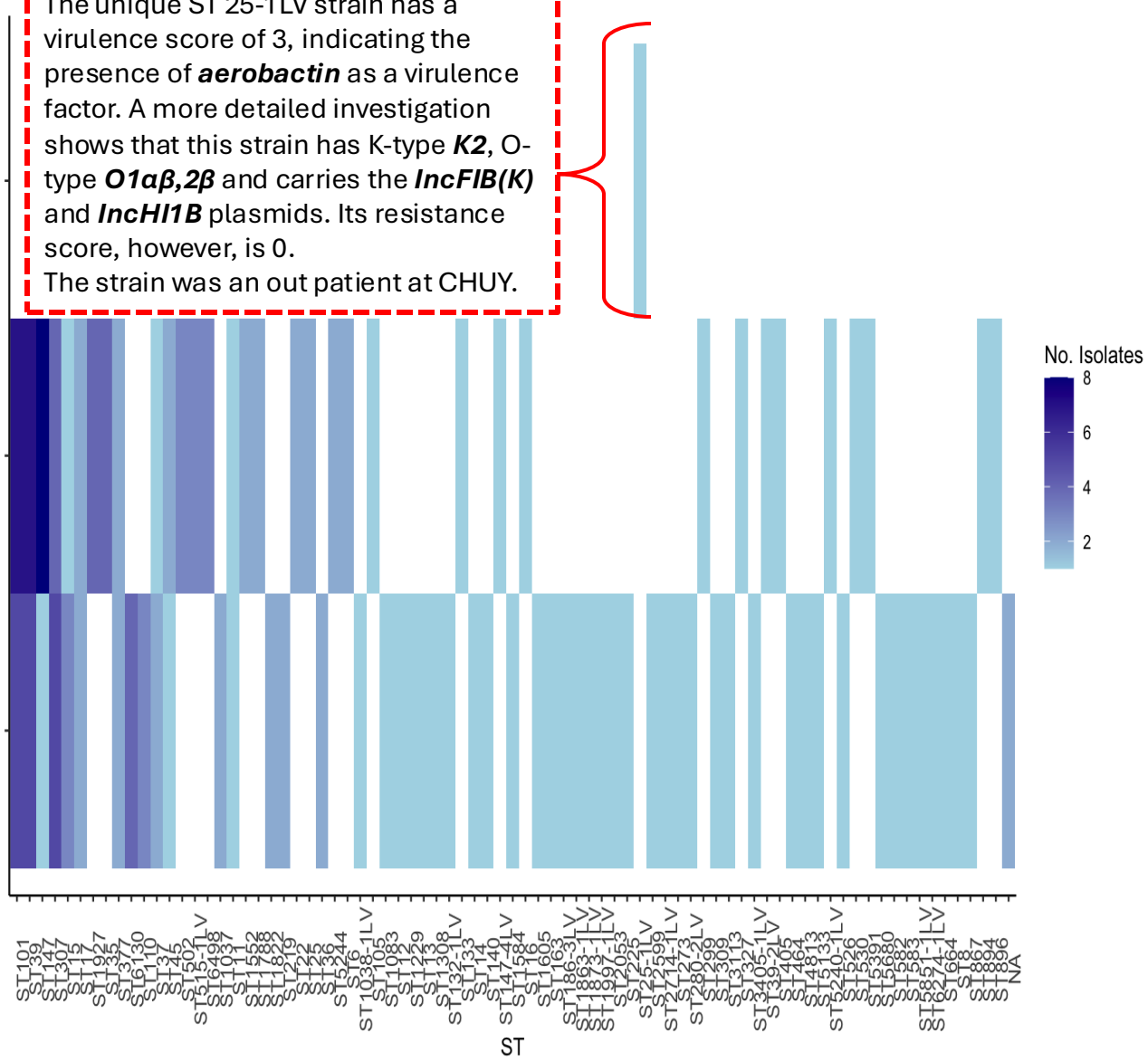
b) Distribution of O type within most prevalent ST

Per ST resistance and virulence score

The unique ST 25-1LV strain has a virulence score of 3, indicating the presence of ***aerobactin*** as a virulence factor. A more detailed investigation shows that this strain has K-type ***K2***, O-type ***O1αβ,2β*** and carries the ***IncFIB(K)*** and ***IncHI1B*** plasmids. Its resistance score, however, is 0.
The strain was an out patient at CHUY.

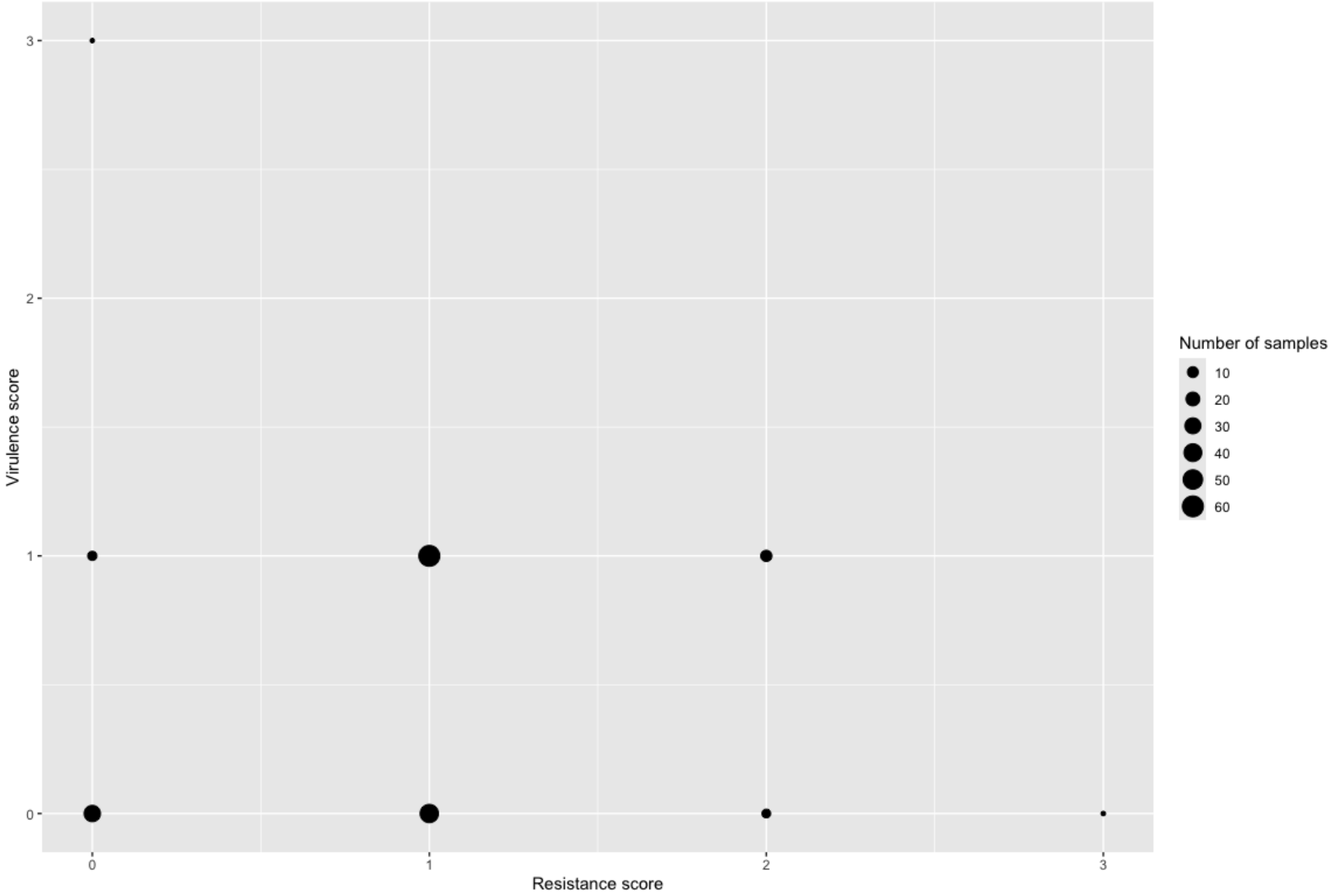


a) Distribution of resistance score per ST

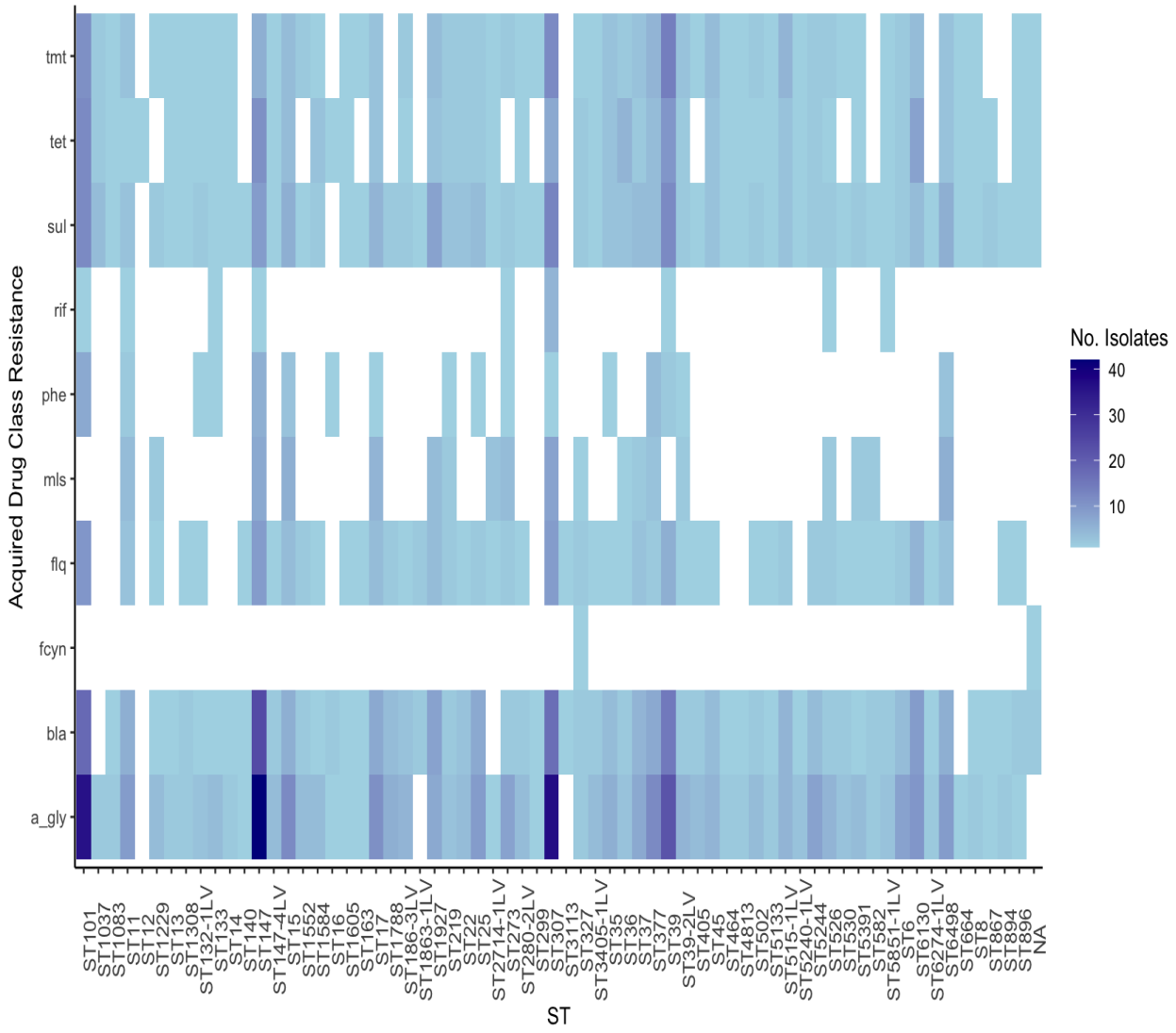


b) Distribution of virulence score per ST

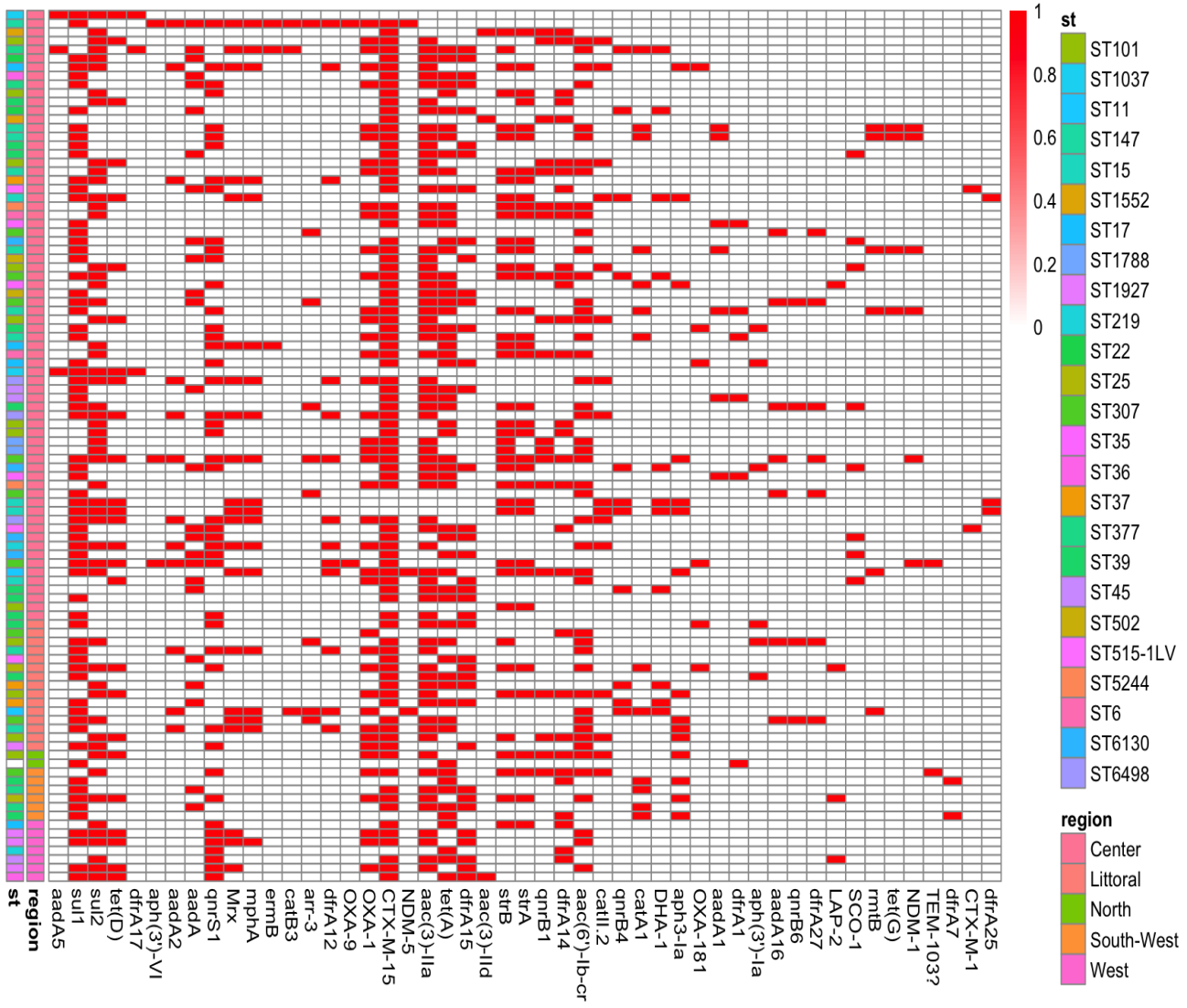
Convergence of resistance and virulence



Most prevalent ST : Resistance Classes and Associated Genes



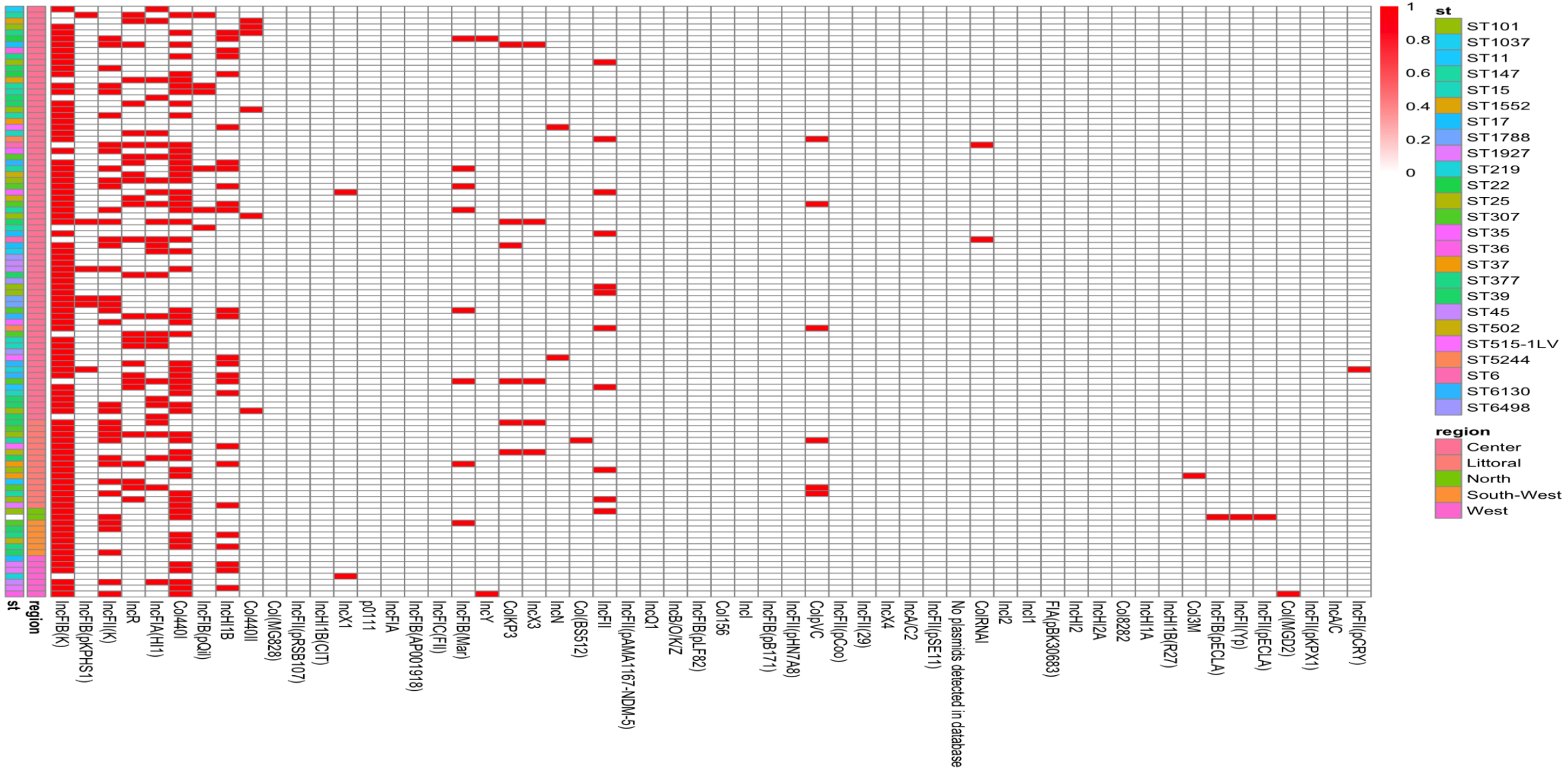
a) Distribution of acquired drug resistance class per ST



b) Distribution of resistance genes per region and 30 most prevalent ST

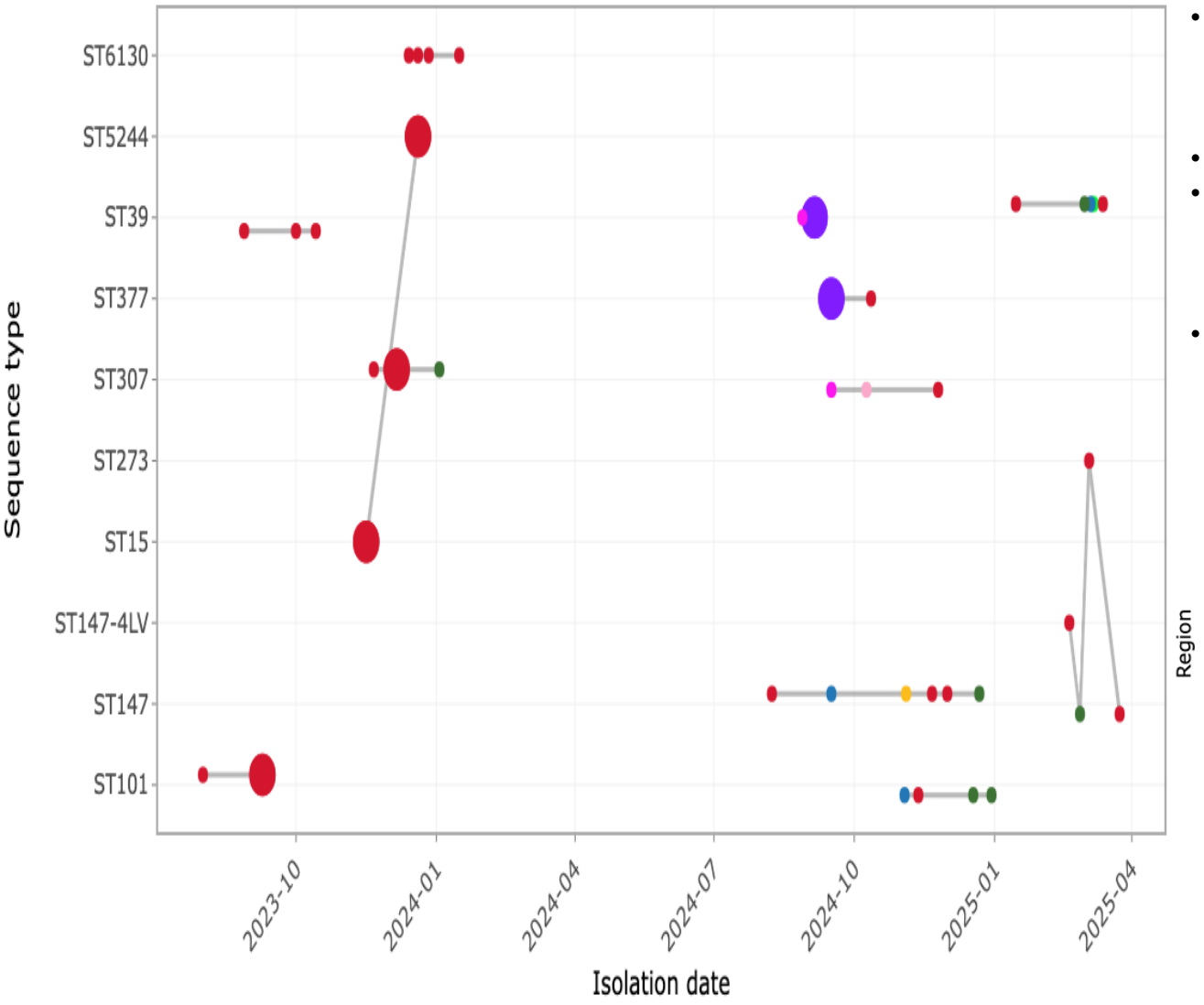
These two plots show MDR in our isolates, with **CTX-M-15**, **aac(3)-IIa**, **sul1** and **tet(A)** being the most prevalent resistance genes across the 30 most common STs

Most prevalent ST : Plasmids distribution



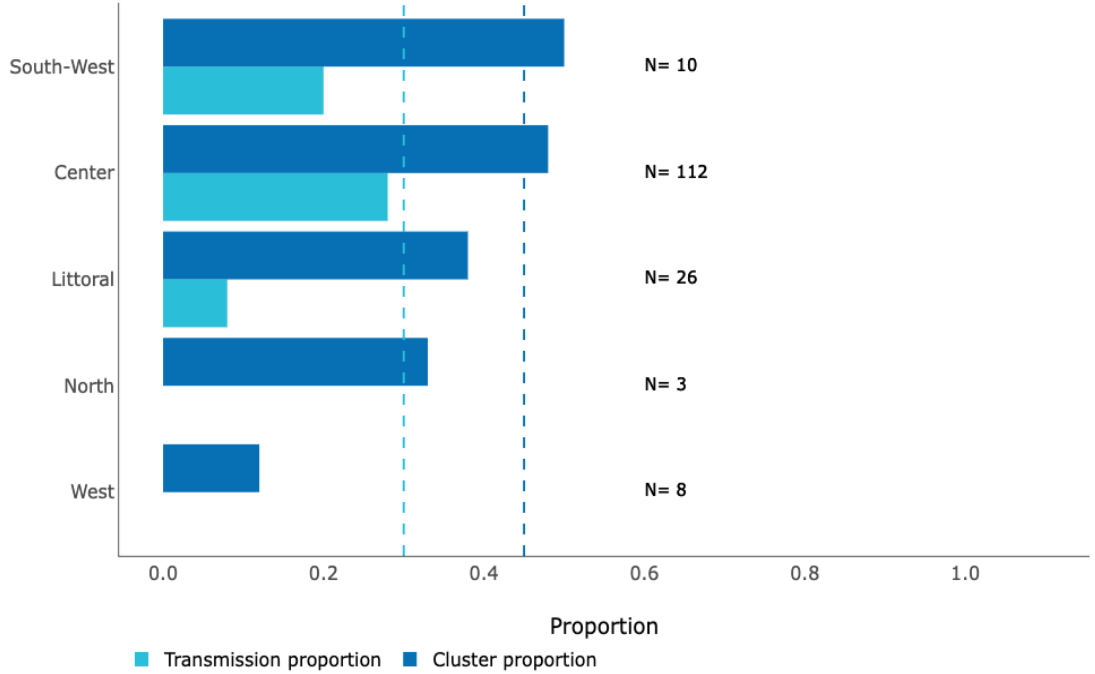
Of the 30 most prevalent STs, only 25 carry plasmids. Among these strains, the majority harbor the ***IncFIB(K)*** plasmid.

Results of transmission analysis



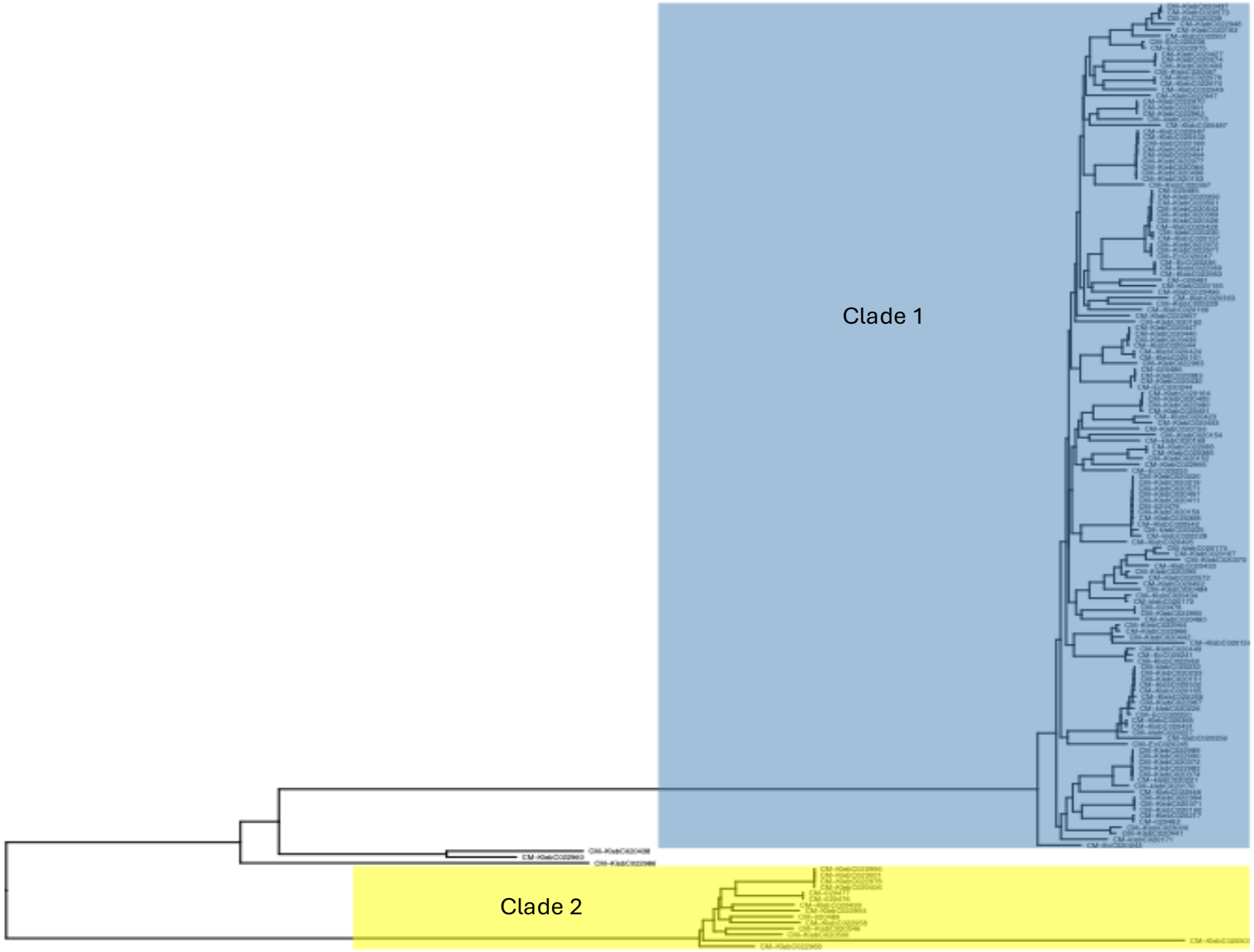
a) Clusters plot per site

- **Using hyperparameters :**
  - Distance threshold = 30
  - Temporal distance threshold (weeks) = 7
  - Spatial Cluster = Country
- There are 12 clusters in our data with a country clustering
- **Observed transmissions:**
  - Between laboratories in the Centre and Littoral regions
  - Between laboratories in the Centre and South-West regions
  - Between laboratories within the Centre region
- The Center region is the one with the most transmission

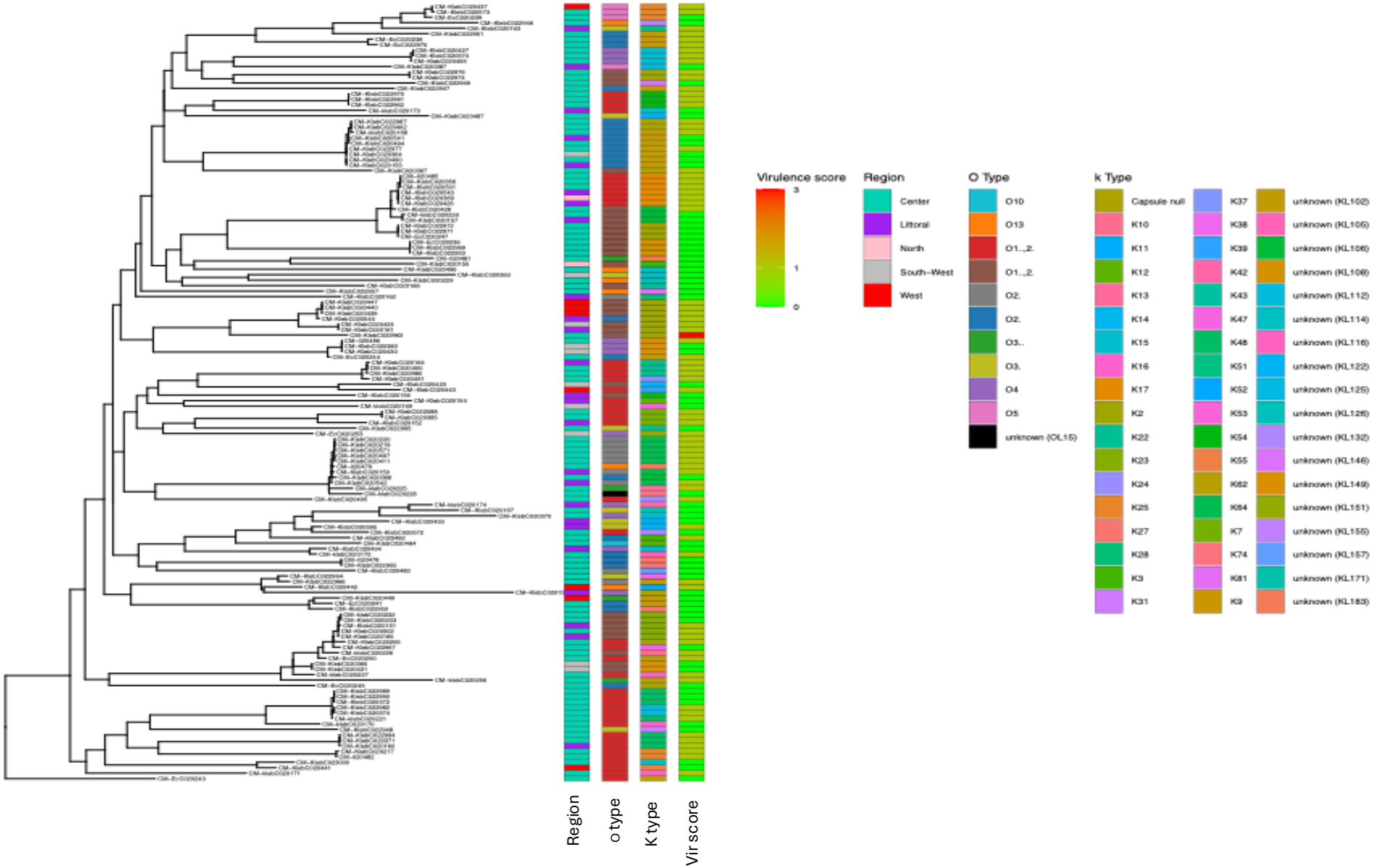


b) Stratified cluster statistics

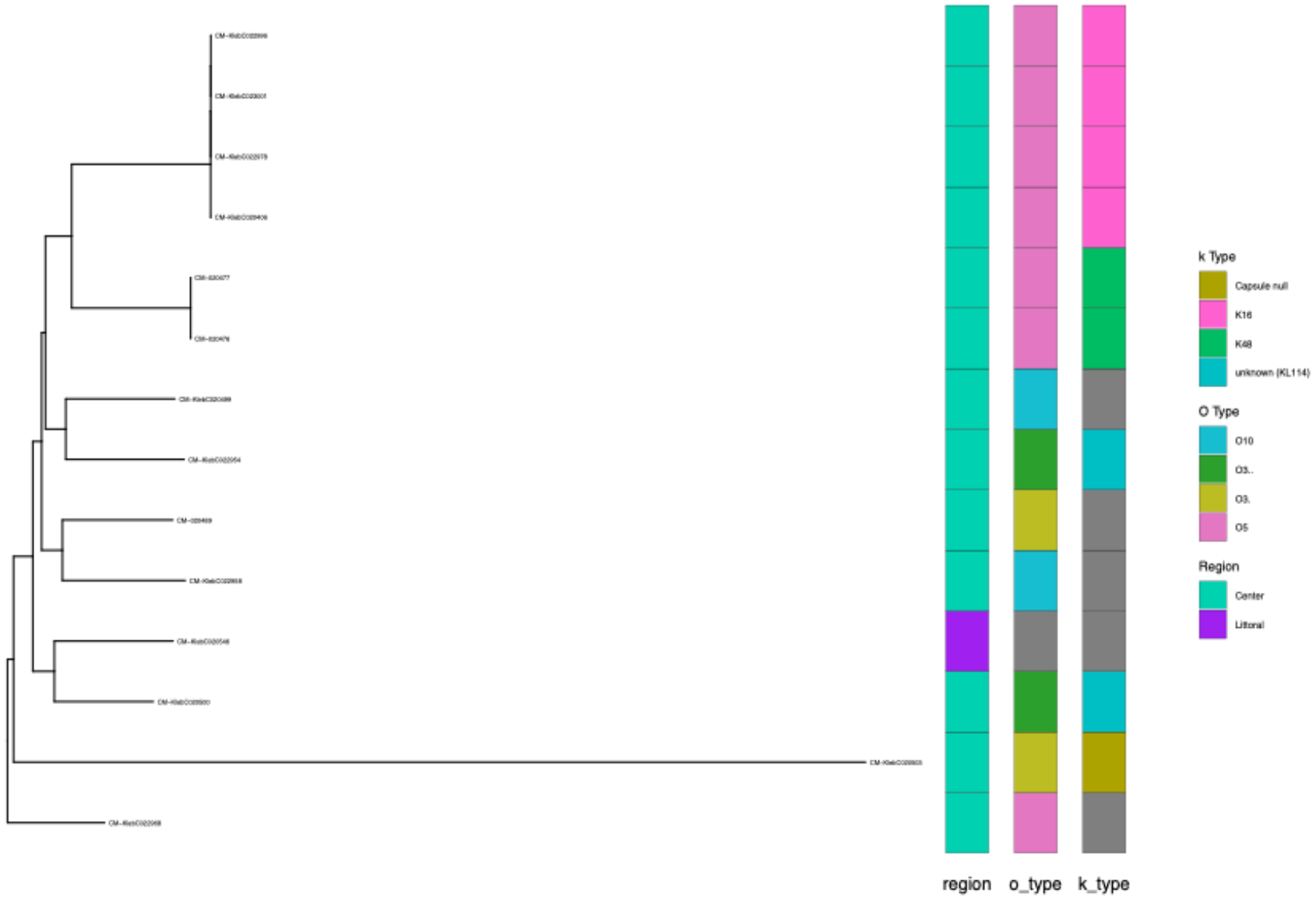
Results of phylogenetic analysis (1/)



Results of phylogenetic analysis : Clade 1 (2/3)



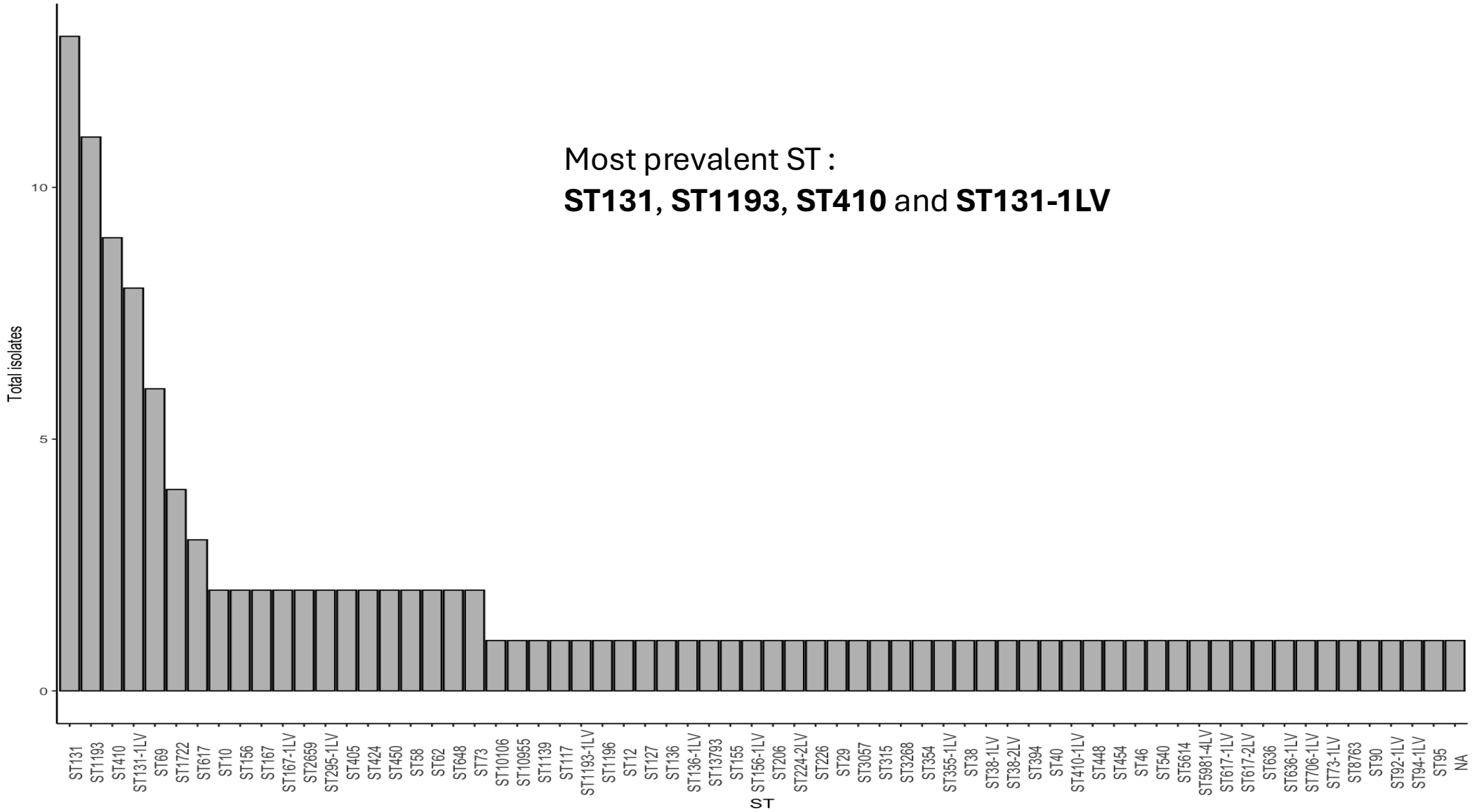
Results of phylogenetic analysis : Clade 2 (3/3)



E. COLI/SHIGELLA ANALYSIS RESULTS

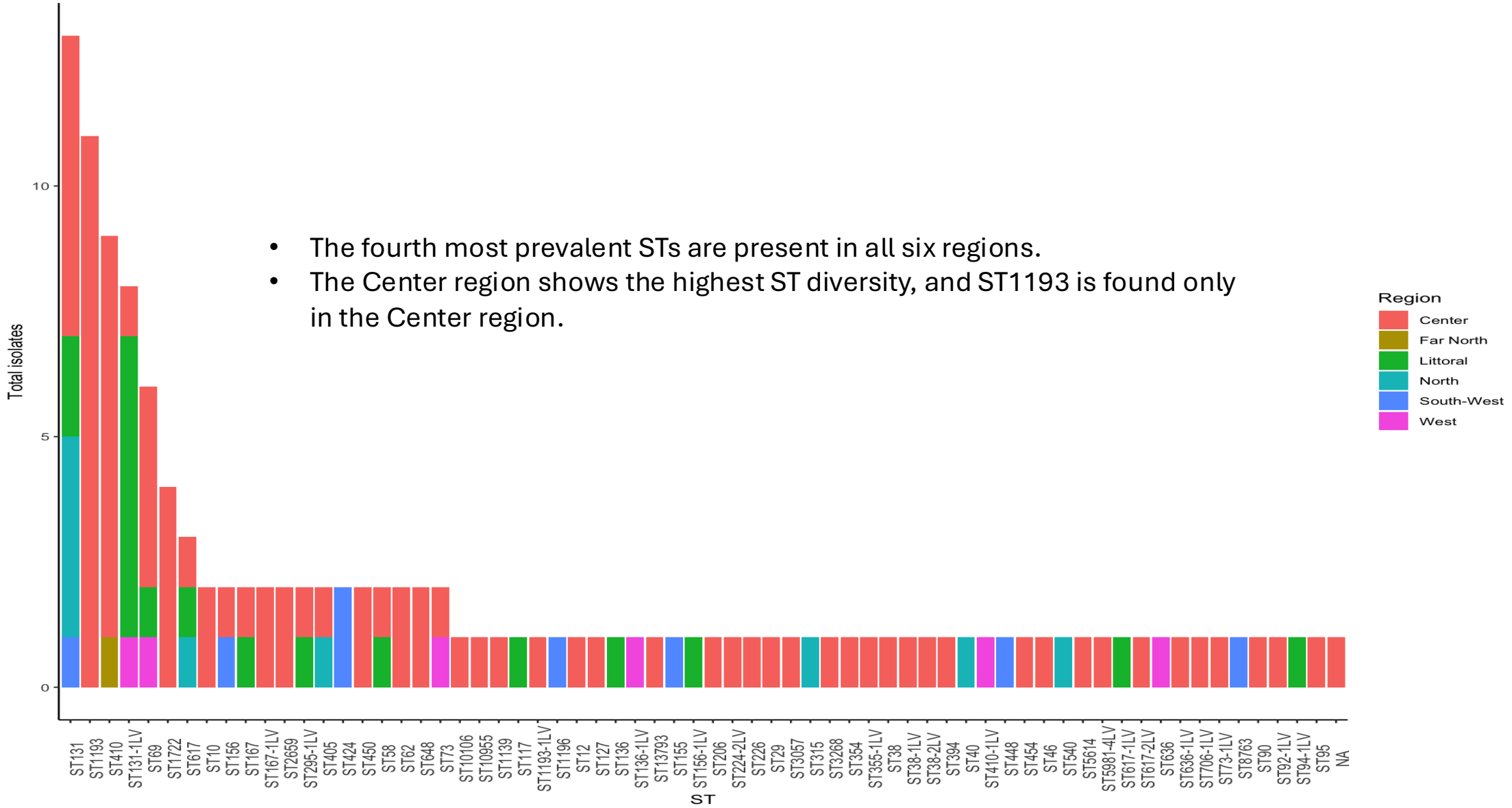
ST prevalence distribution

Most prevalent ST :
ST131, **ST1193**, **ST410** and **ST131-1LV**

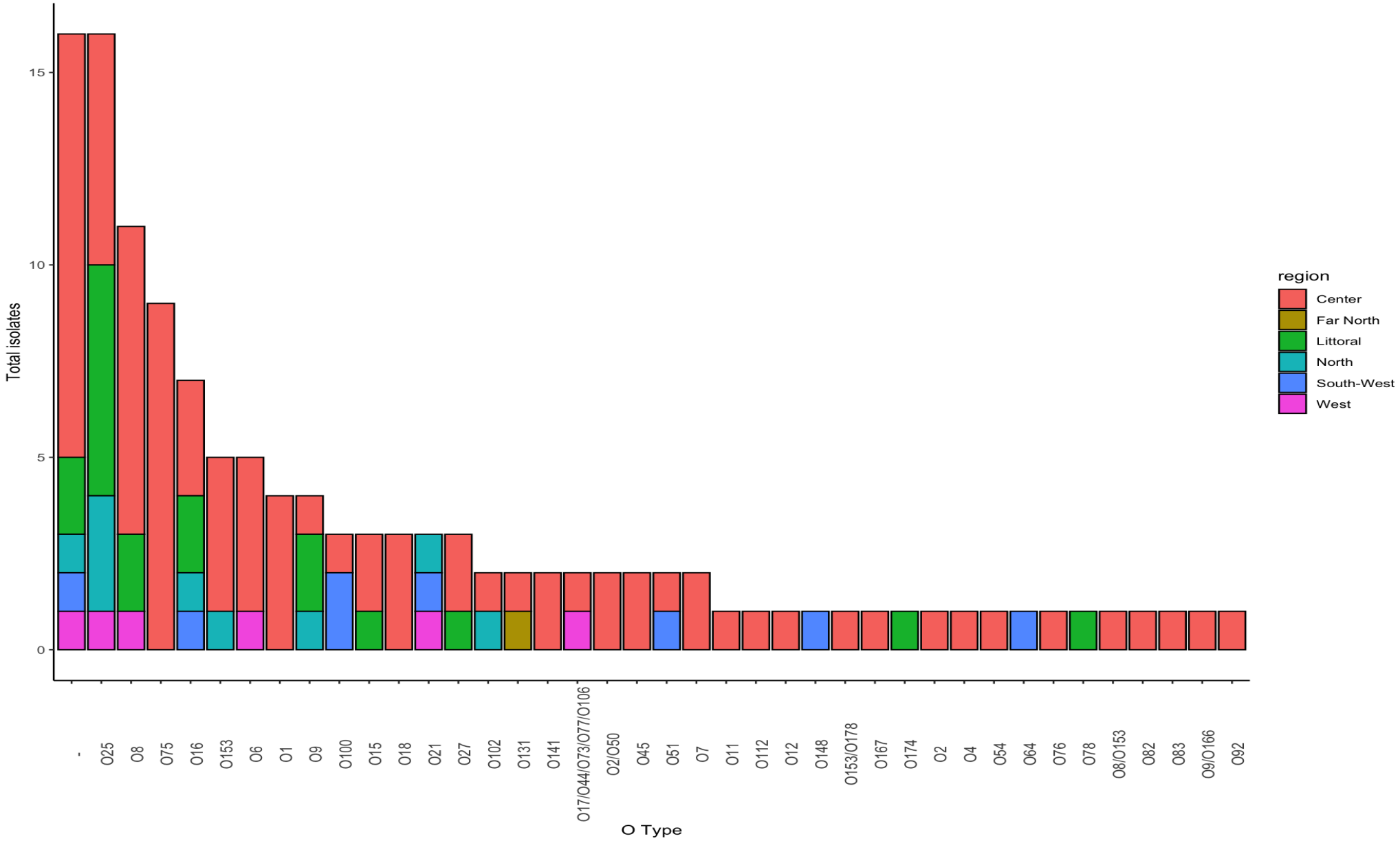


ST prevalence distribution per region

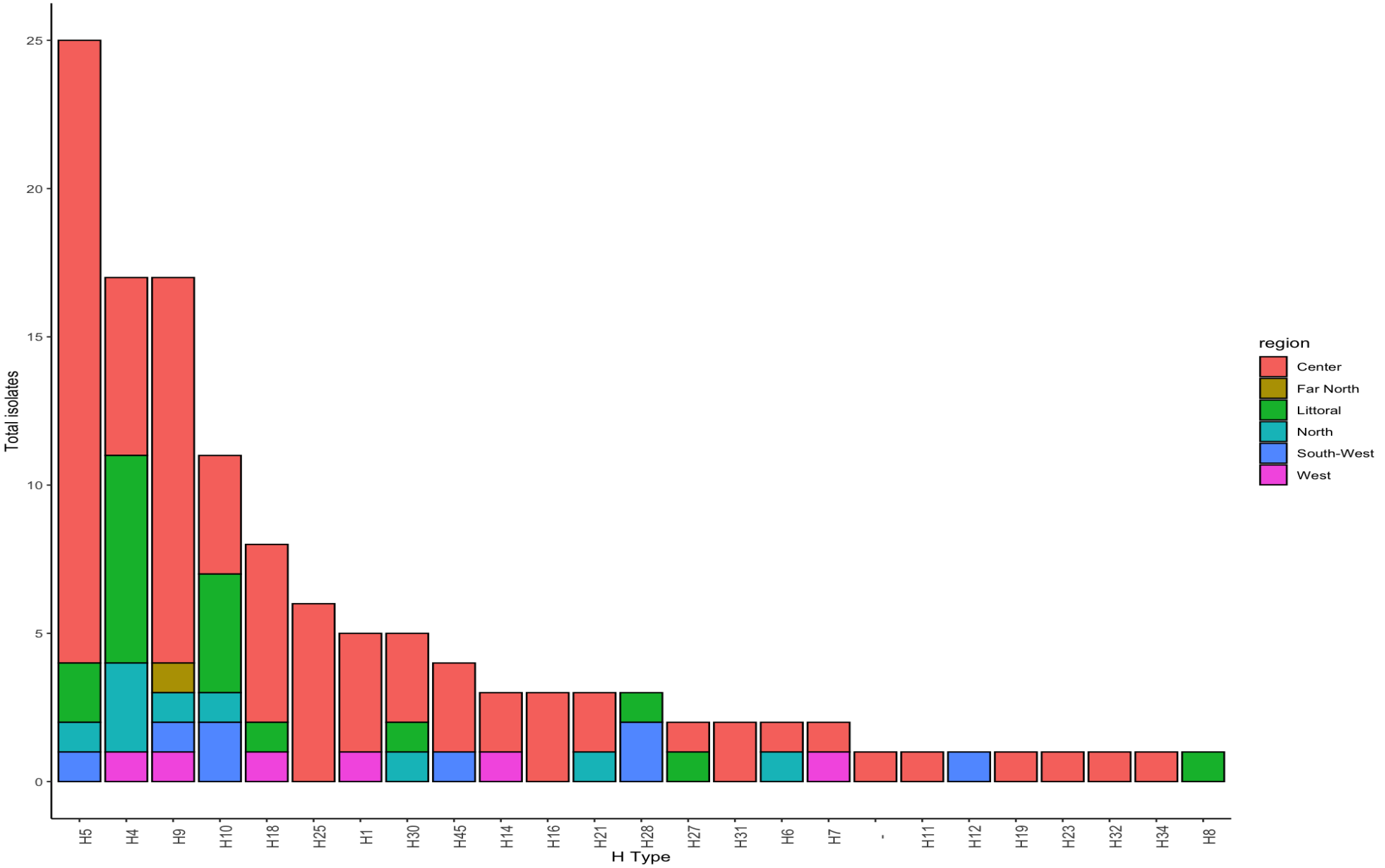
- The fourth most prevalent STs are present in all six regions.
- The Center region shows the highest ST diversity, and ST1193 is found only in the Center region.



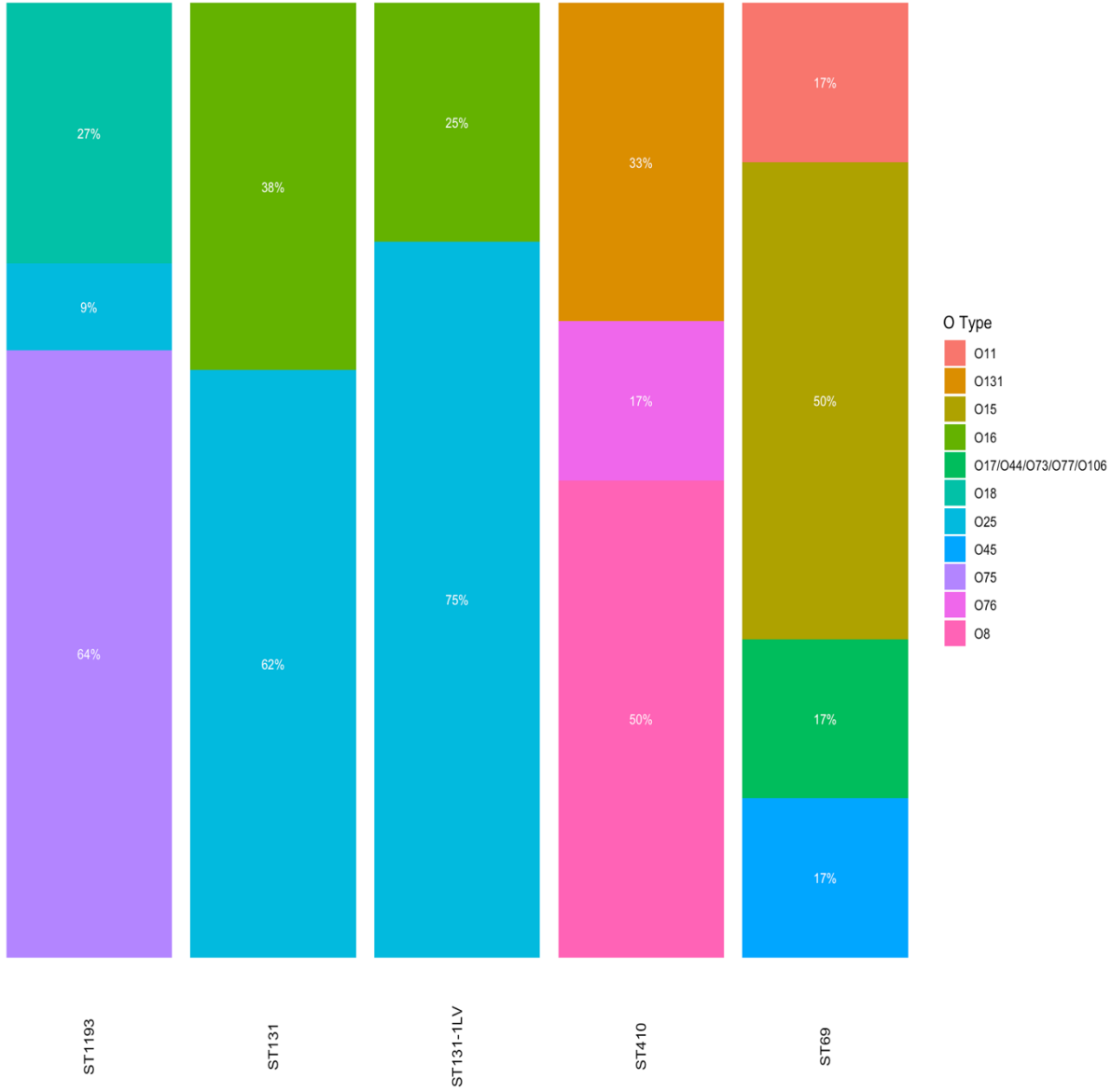
O type Distribution per region



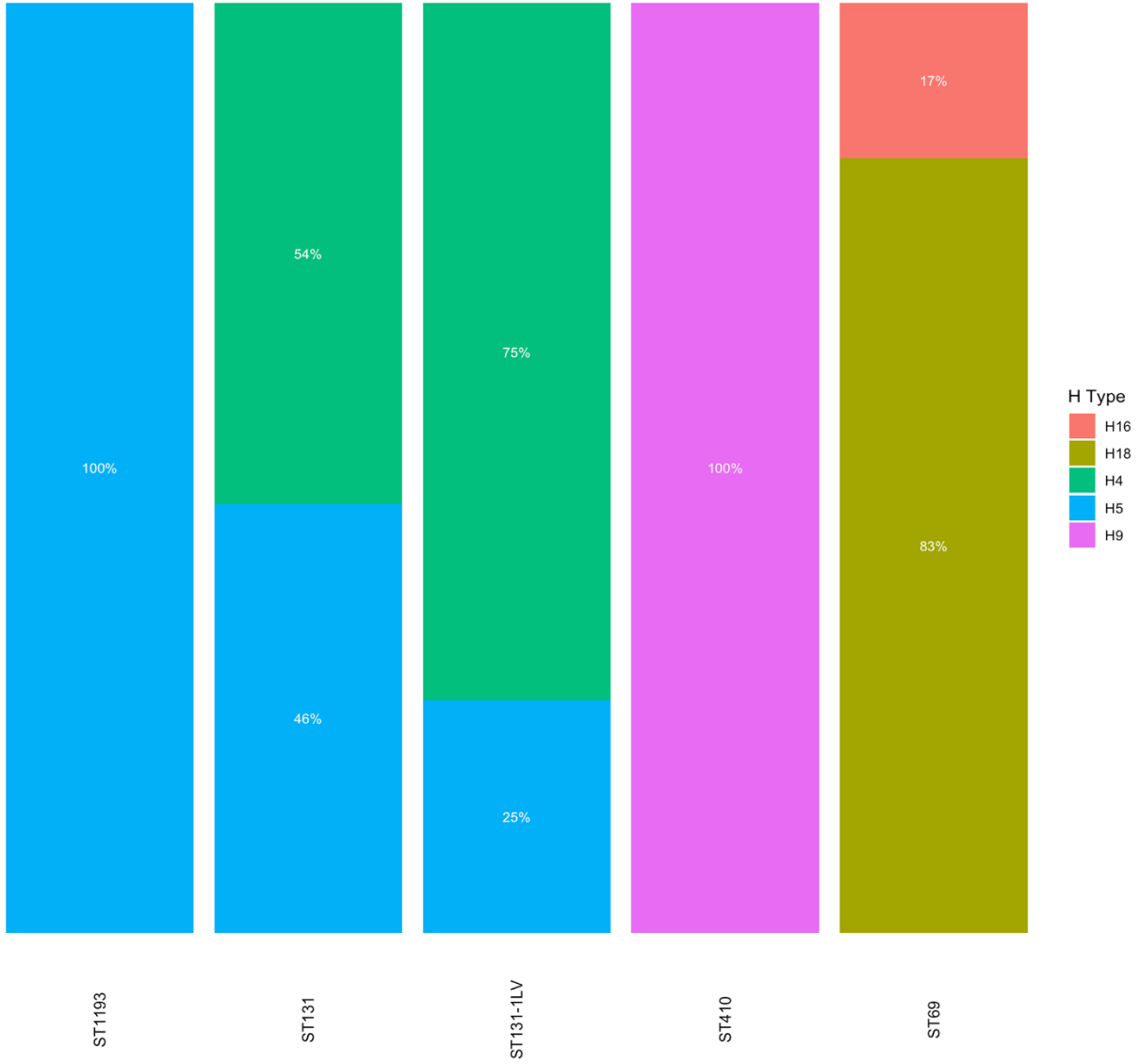
H type Distribution per region



5 most prevalent ST : O type and H type distribution

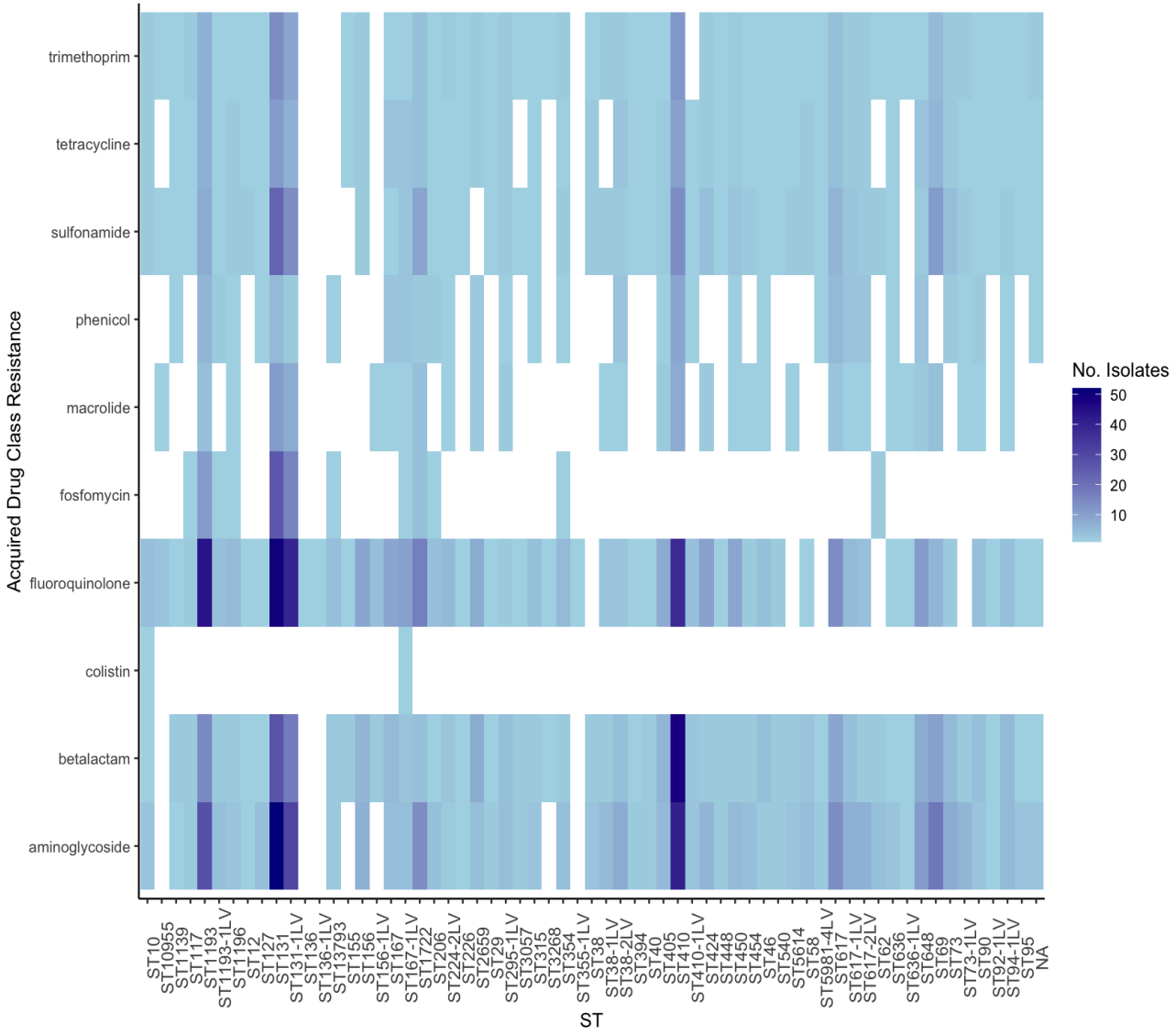


a) Distribution of O type within most prevalent ST

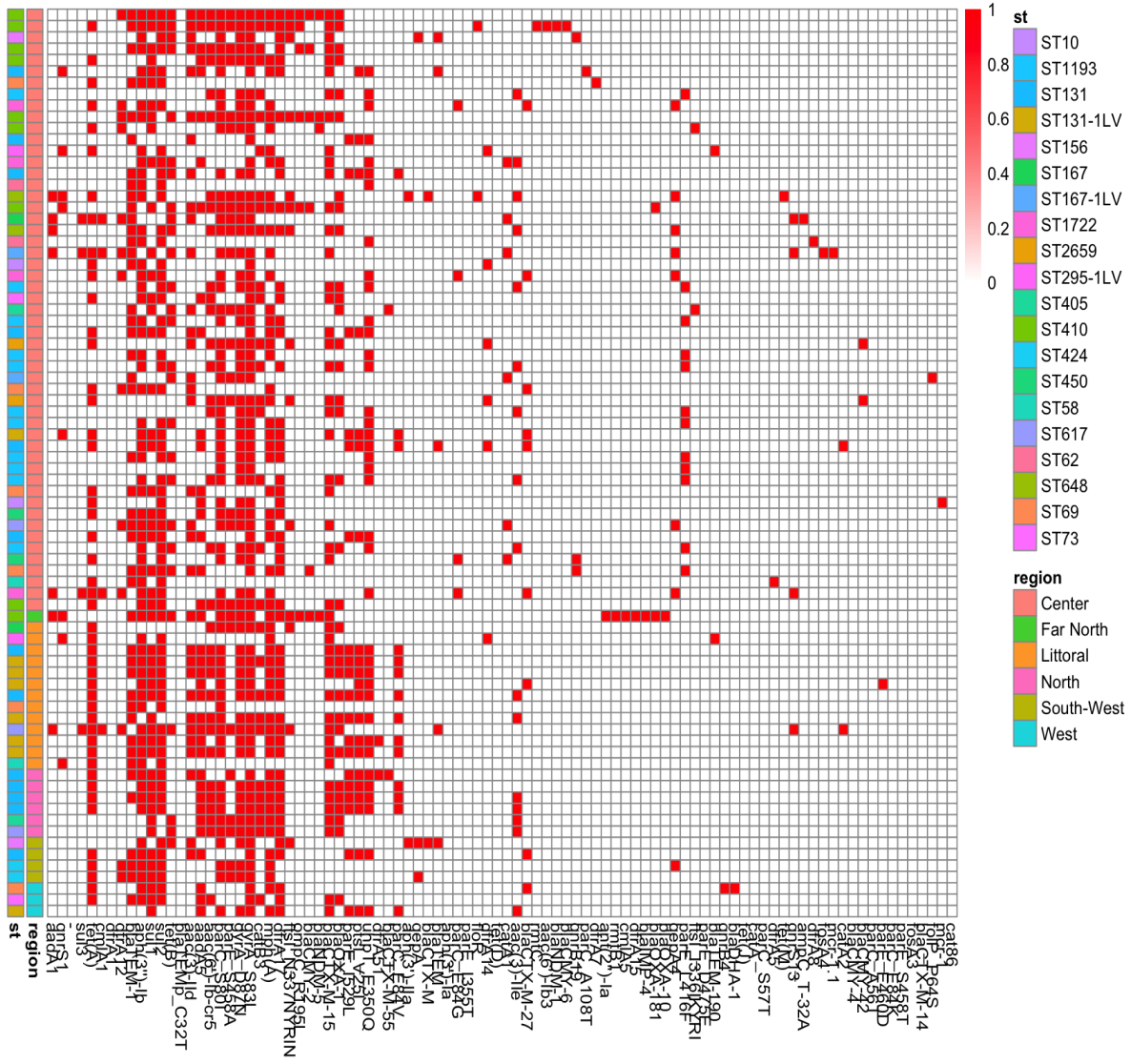


b) Distribution of H type within most prevalent ST

Most prevalent ST : Resistance Classes and Associated Genes

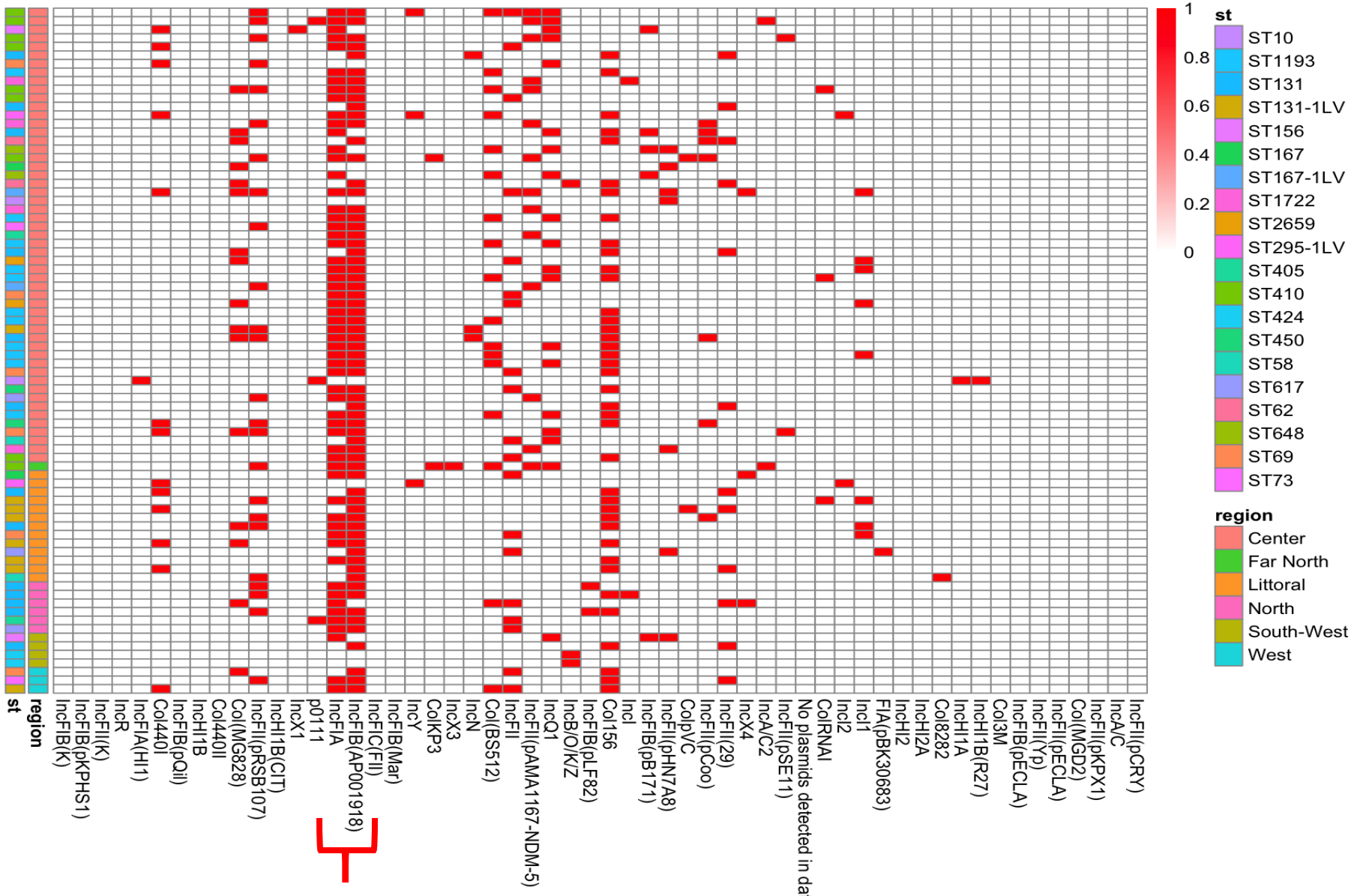


a) Distribution of acquired drug resistance class per ST

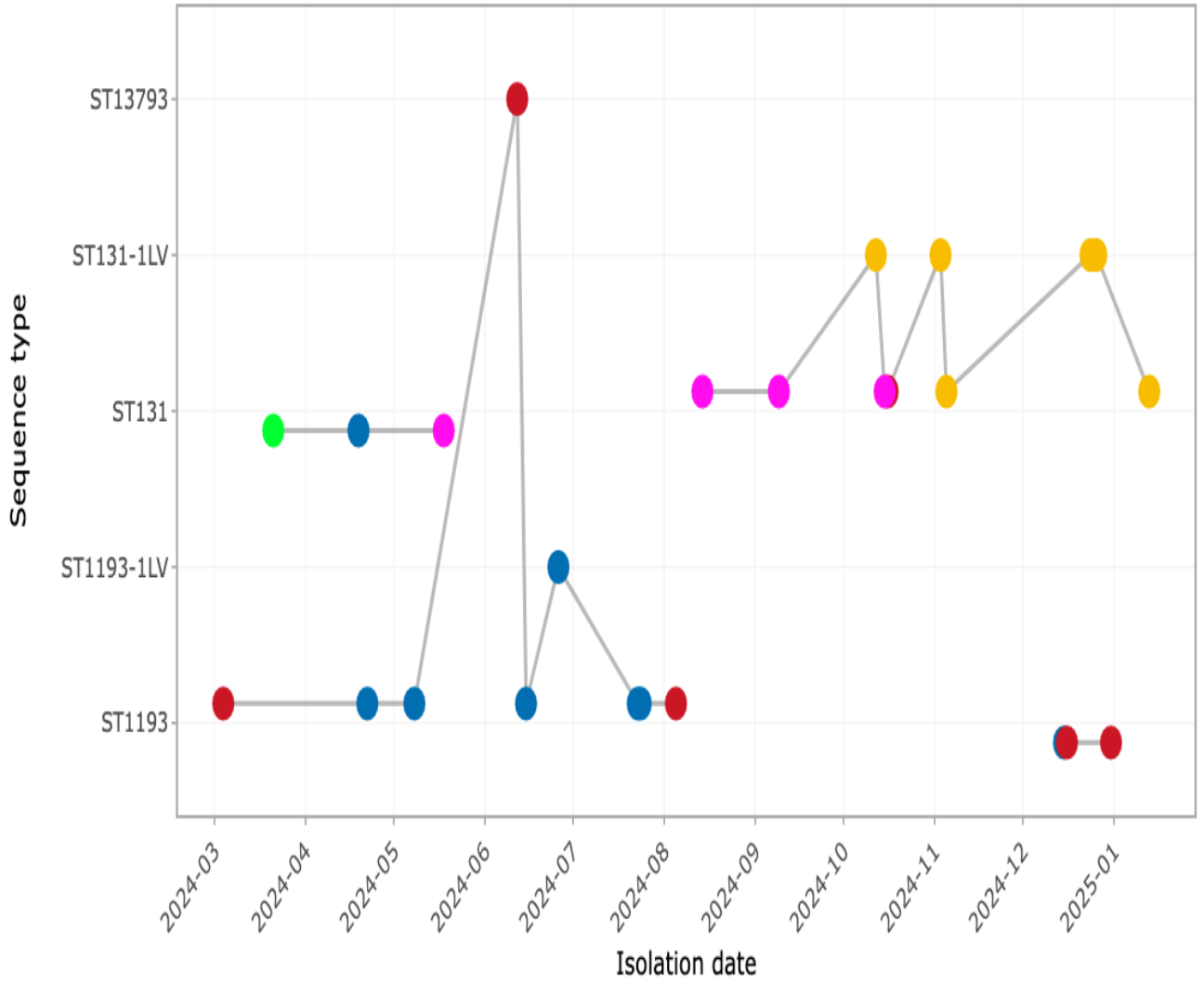


b) Distribution of resistance genes per region and 20 most prevalent ST

Most prevalent ST : Plasmids distribution

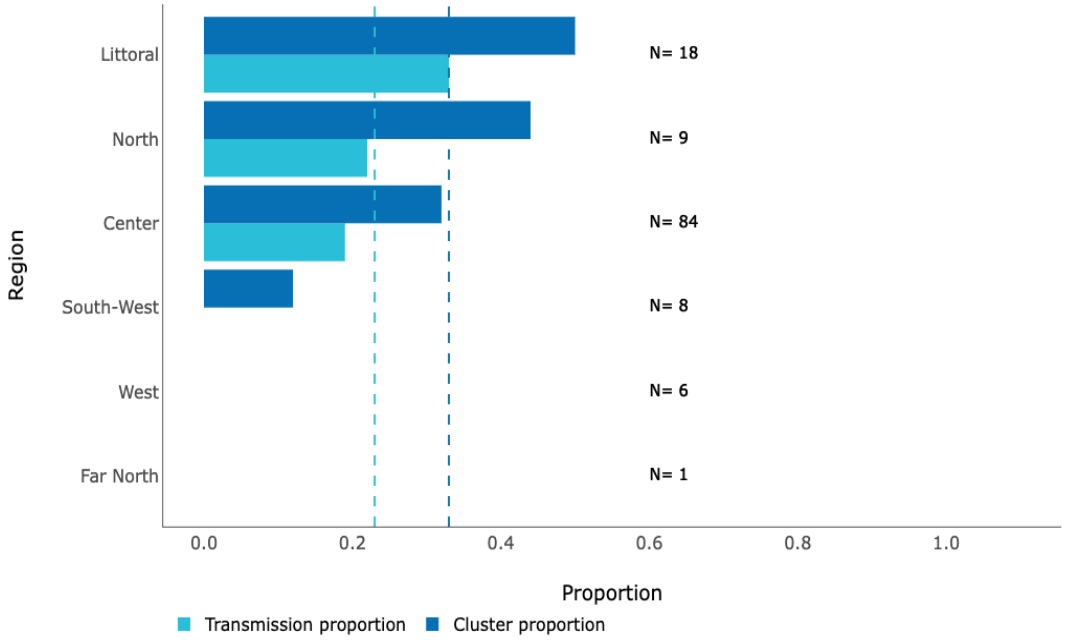


Results of transmission analysis



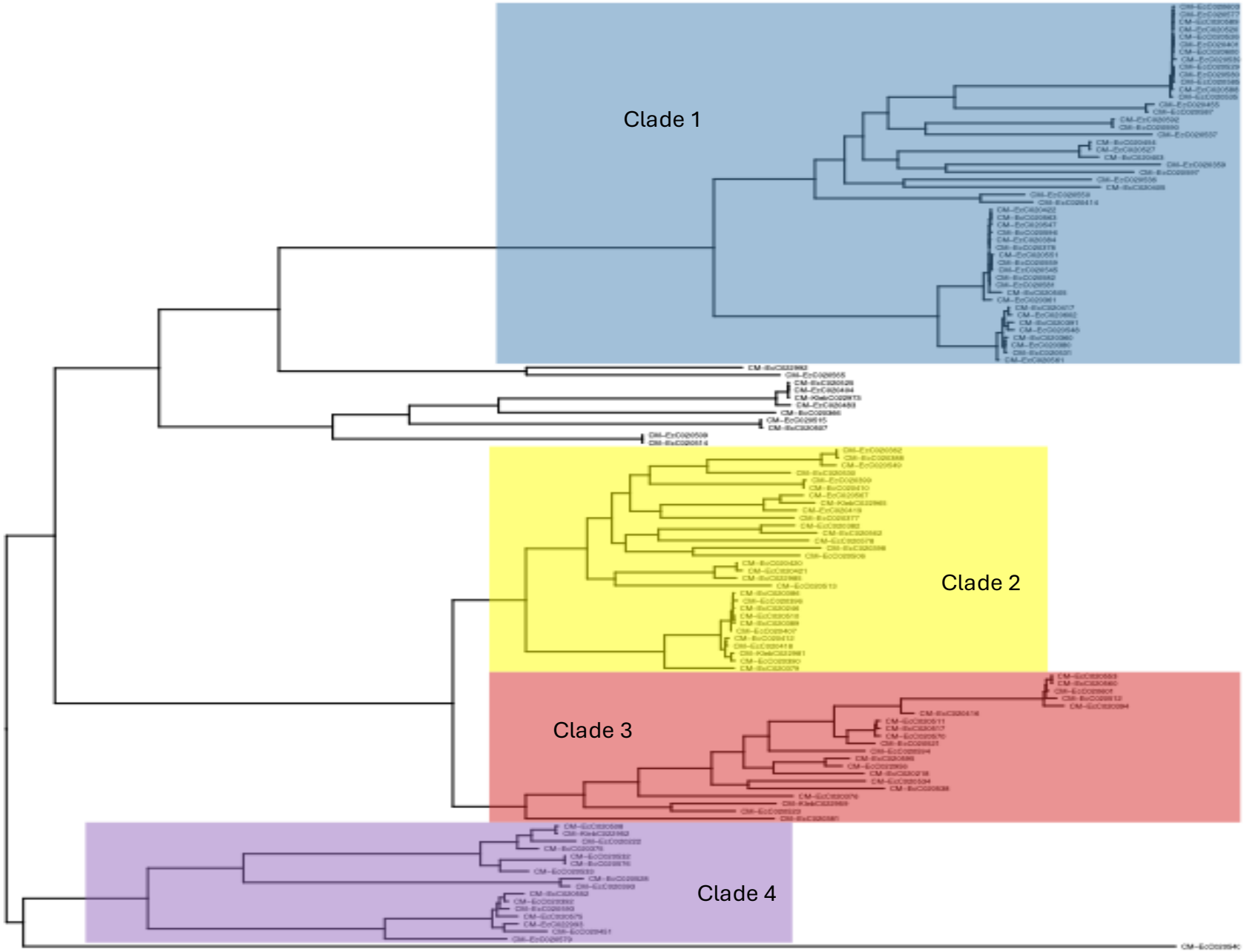
a) Clusters plot per site

- **Using hyperparameters :**
  - Distance threshold = 30
  - Temporal distance threshold (weeks) = 4
  - Spatial Cluster = Country
- There are 4 clusters in our data with a country clustering
- **Observed transmissions:**
  - Between laboratories in the Centre, North and Littoral regions
  - Between laboratories in the Centre and North regions
  - Between laboratories within the Centre region
- The Littoral region is the one with the most transmission

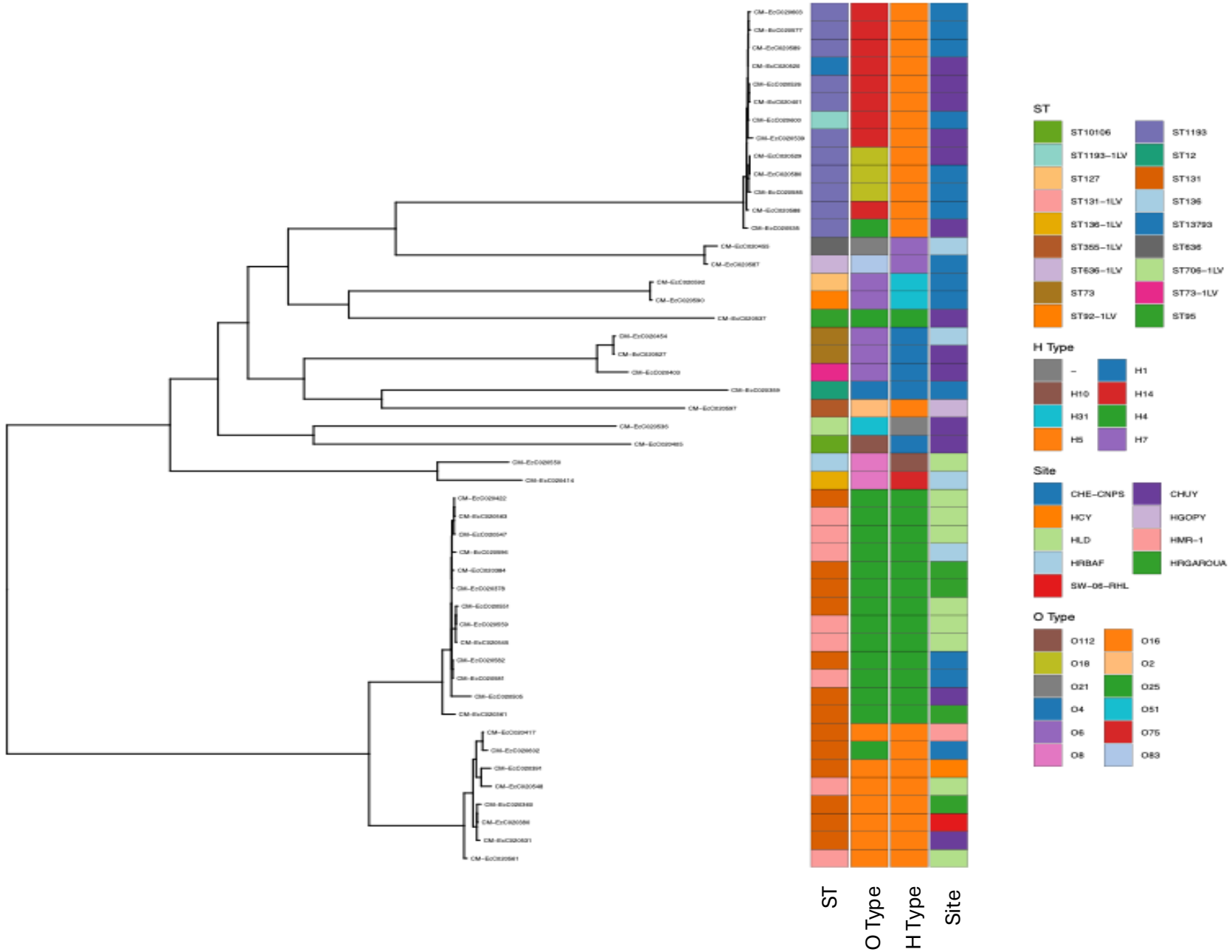


b) Stratified cluster statistics

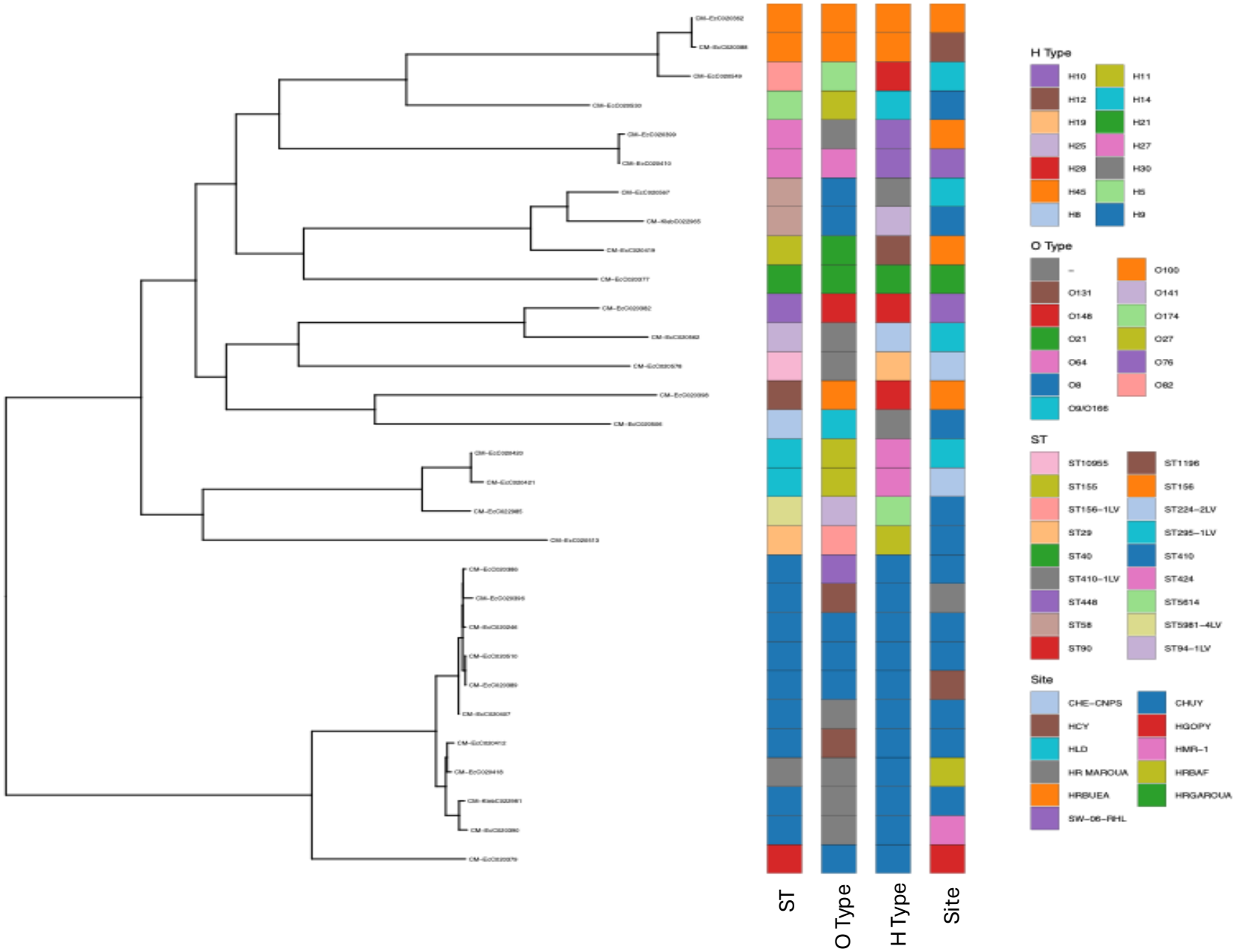
Results of phylogenetic analysis (1/5)



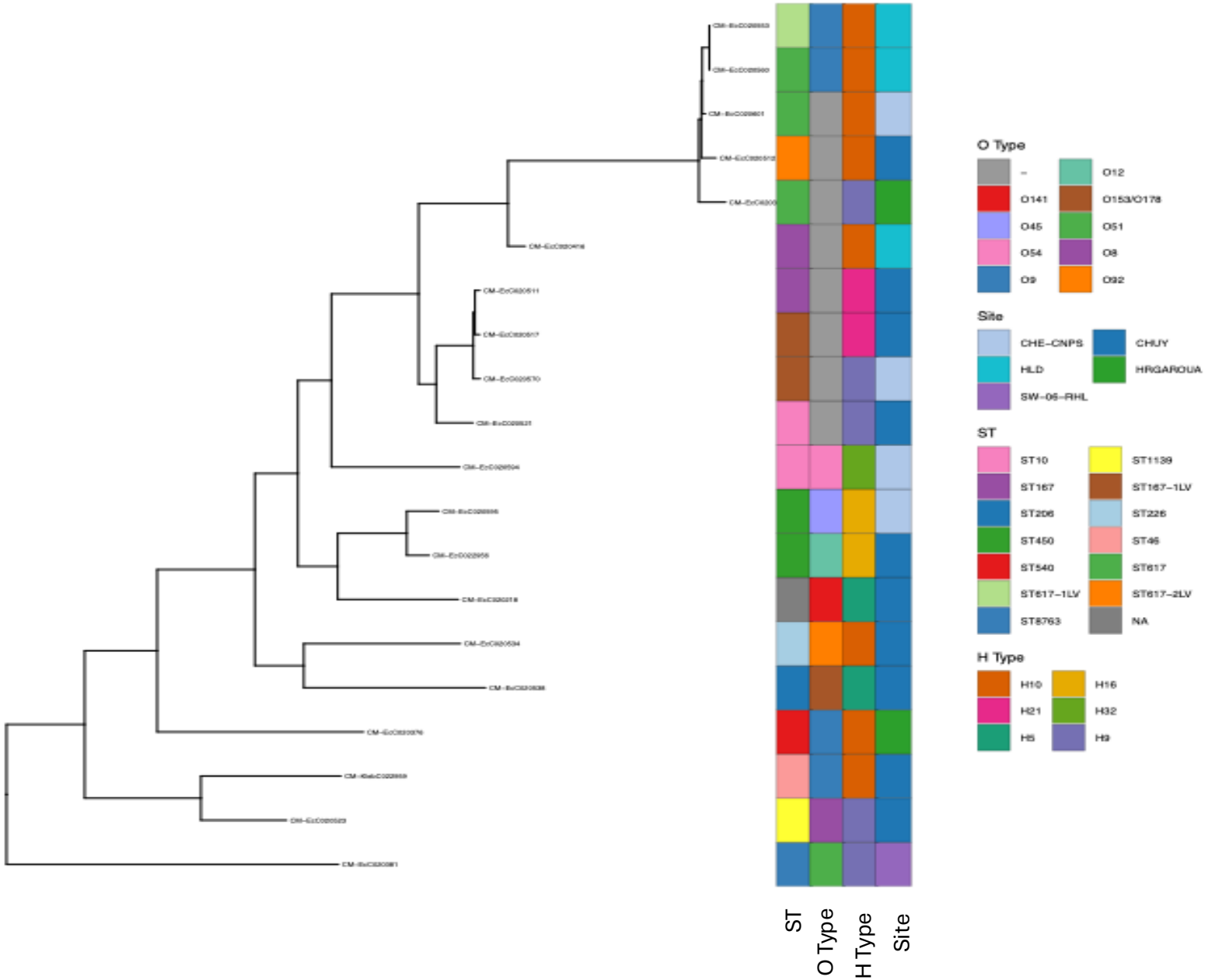
Results of phylogenetic analysis : Clade 1 (2/5)



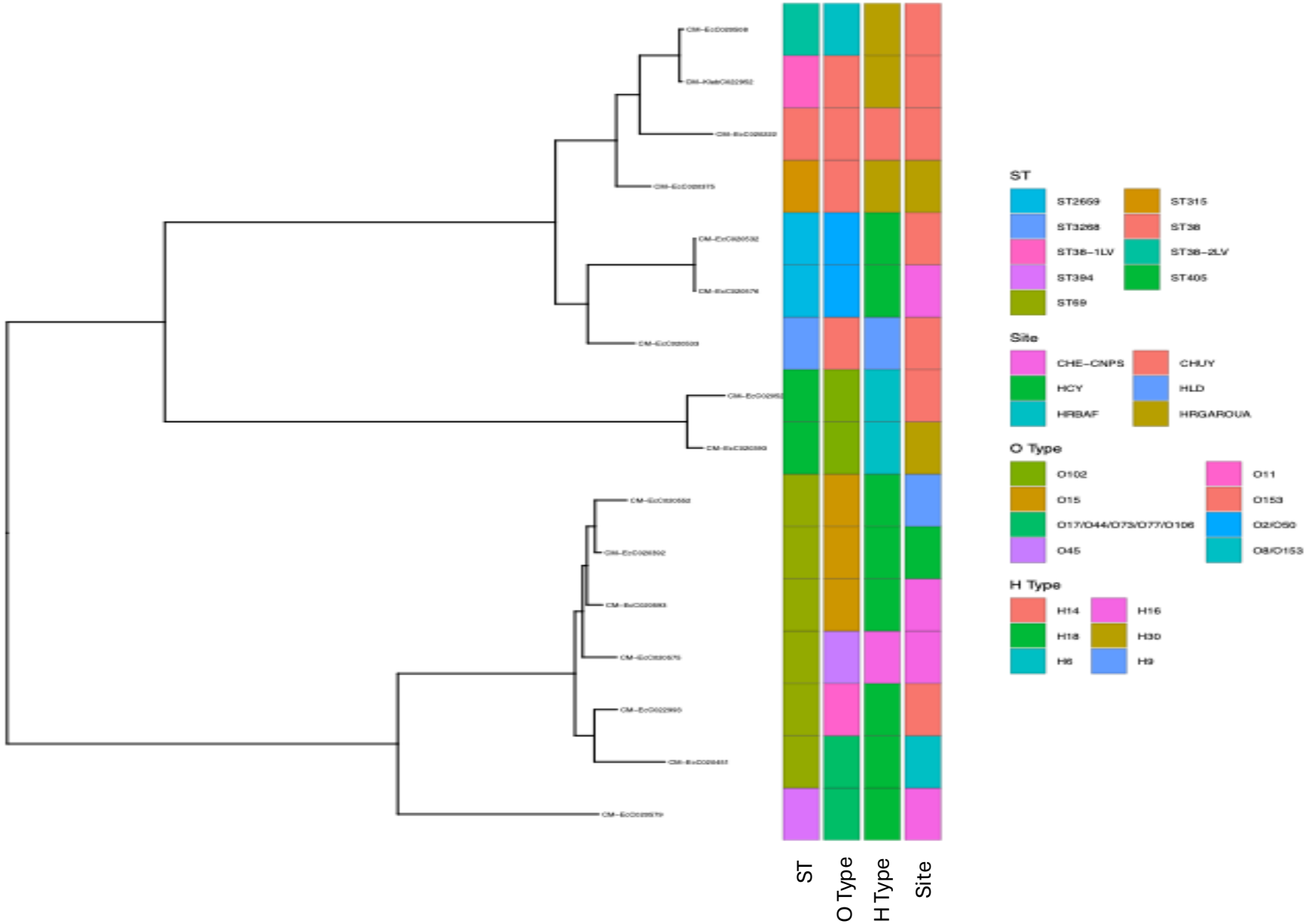
Results of phylogenetic analysis : Clade 2 (3/5)

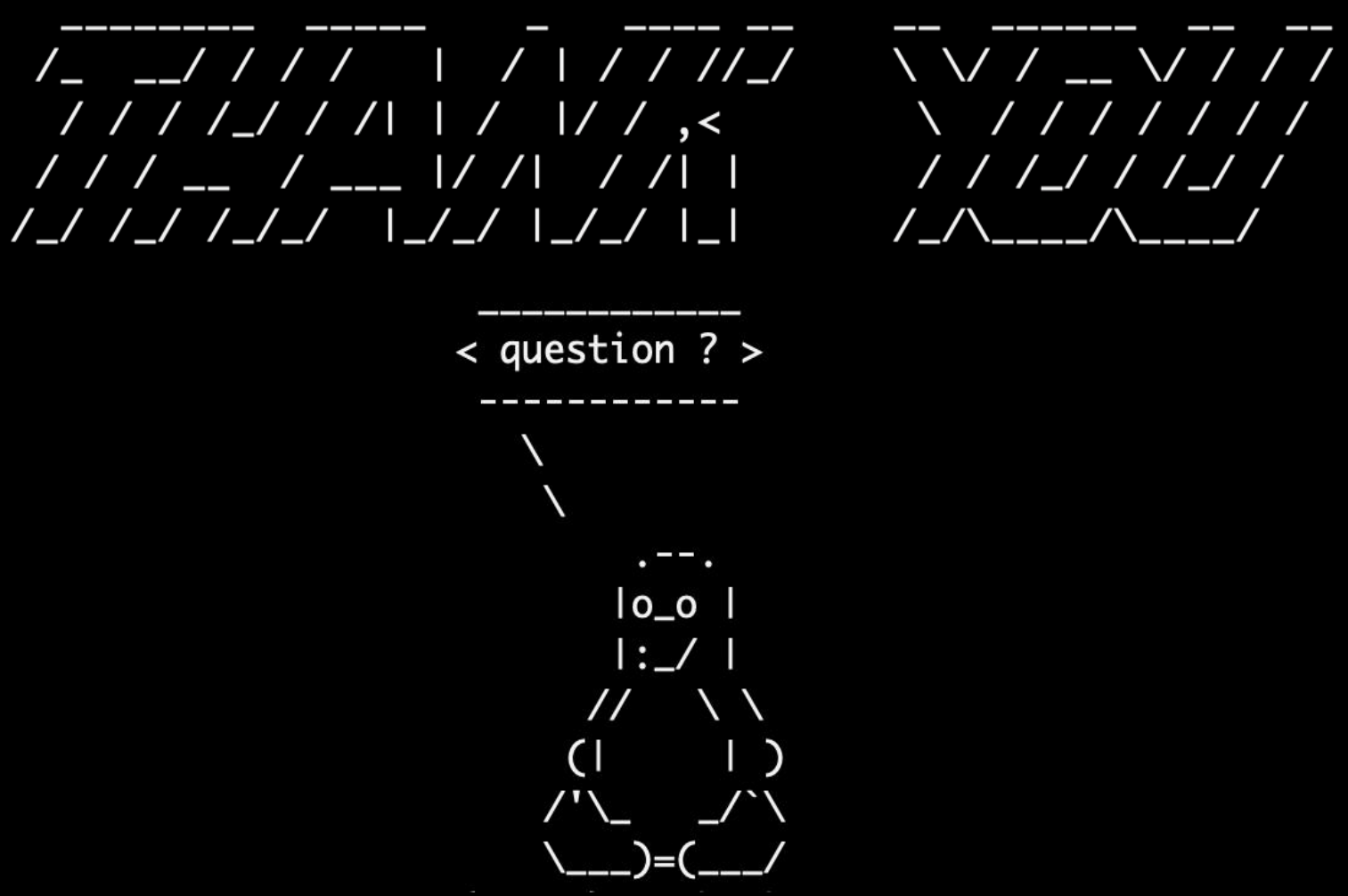


Results of phylogenetic analysis : Clade 3 (4/5)



Results of phylogenetic analysis : Clade 4 (5/5)





By MMCJ