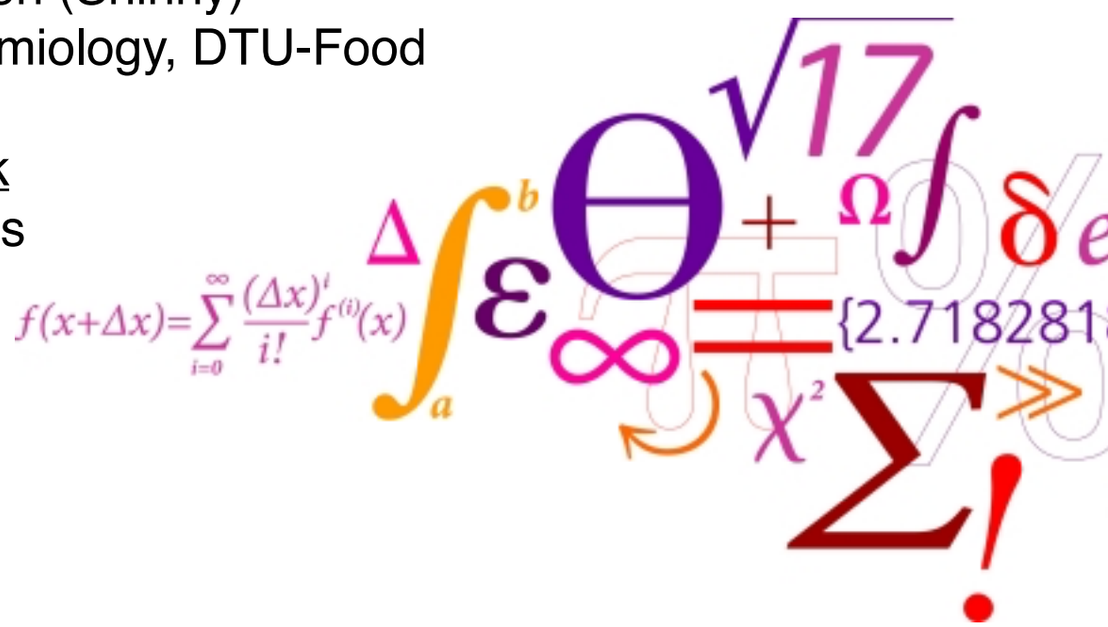
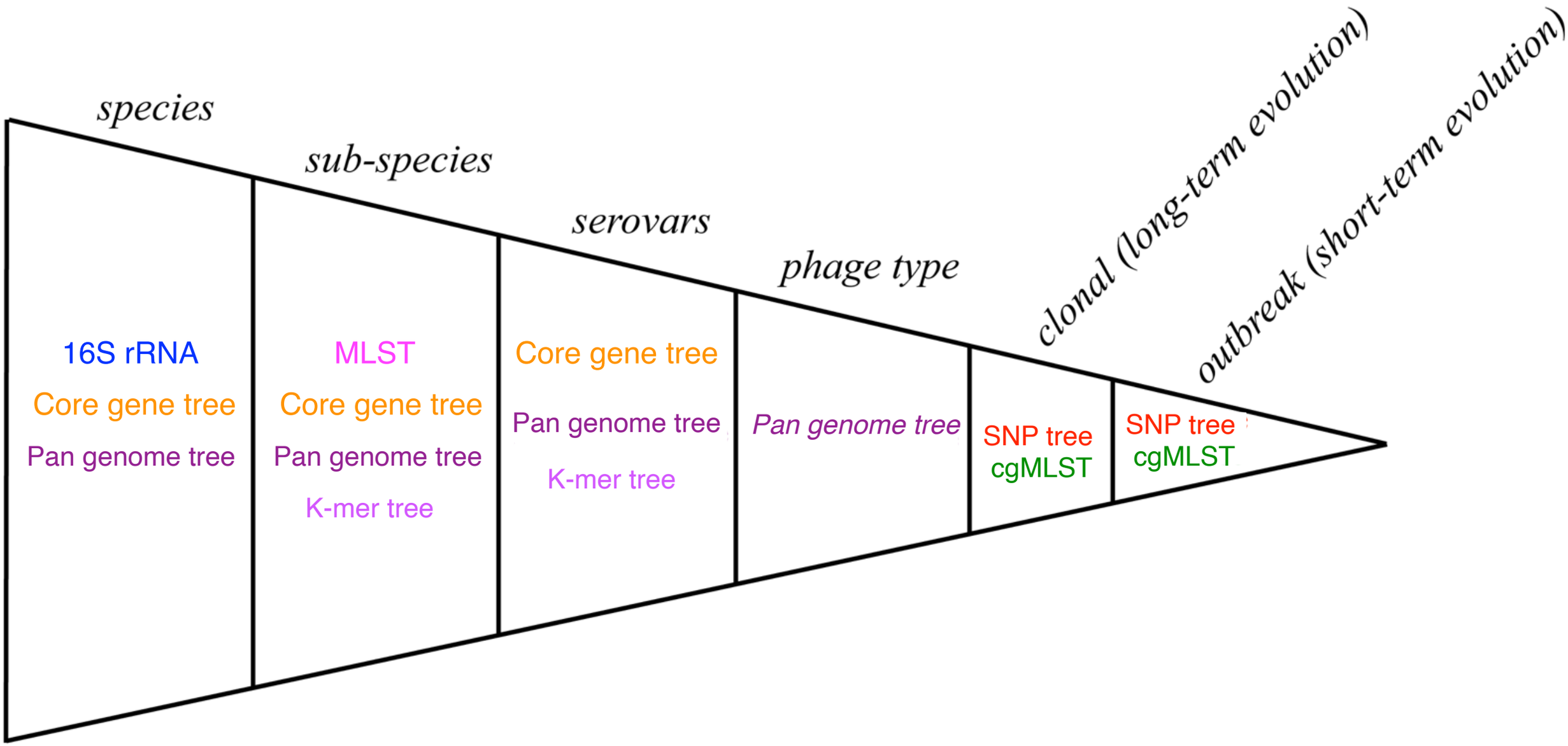
Phylogeny for outbreak investigation using cgMLST/wgMLST tree

Pimlapas Leekitcharoenphon (Shinny)
Research Group of Genomic Epidemiology, DTU-Food

pile@food.dtu.dk
@ShinnyPimlapas



Infectious Disease Bioinformatics Course, March 2022


species
sub-species
serovars
phage type
clonal (long-term evolution)
outbreak (short-term evolution)
16S rRNA
Core gene tree
Pan genome tree
MLST
Core gene tree
Pan genome tree
K-mer tree
Core gene tree
Pan genome tree
K-mer tree
Pan genome tree
SNP tree
cgMLST
SNP tree
cgMLST

Whole genome based phylogeny

- Single nucleotide polymorphism (SNPs) approach
  - Require reference genome
- Gene by gene approach
  - cgMLST, wgMLST
  - No reference genome required
  - Required species specific cgMLST scheme

cgMLST tree

Raw reads

KMA (K-mer alignment)

cgMLST

Best-matched allele numbers

python

Allele number matrix

‘gower’ distance method in R

Pairwise distance

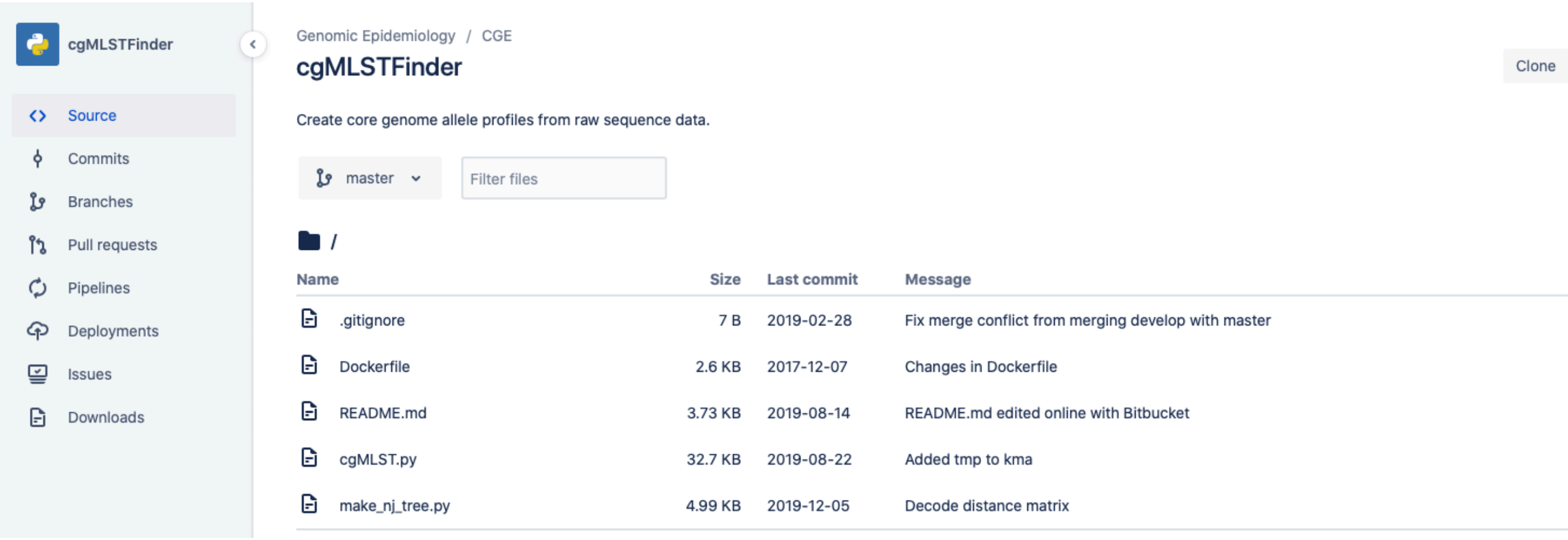
hierarchical clustering in R

cgMLST tree

Allele number matrix

Genome	CAMP0001	CAMP0002	CAMP0003	CAMP0006	CAMP0007	CAMP0009	CAMP0010	CAMP0012	CAMP0013	CAMP0015	CAMP0016	CAMP0017	CAMP0018	CAMP0021	CAMP0022
14035391_S3_L001	724	139	202	80	118	81	18	1	363	1	1	1	1	2	96
14035392_S4_L001	1	30	29	1	1	1	1	1	1	1	1	14	1	22	22
14036372-CAM_S93_L001	17	148	21	20	1040	18	18	18	211	17	4	20	1	15	13
14036375-CAM_S19_L001	30	190	277	23	71	1	1	1	1	31	1	32	1	17	32
14038468_S7_L001	573	480	43	159	679	357	35	2	244	181	4	41	1	2	73
14040169_S24_L001	140	124	170	102	168	67	13	77	64	53	33	37	1	41	55
14040195_S28_L001	55	57	60	2	65	22	22	22	225	221	14	49	1	127	21
14040713-CAM_S94_L001	111	111	72	51	81	58	28	31	36	251	25	37	1	33	40
14044103_S26_L001	107	107	29	26	59	48	35	2	1	1	2	46	1	2	2
14044105_S28_L001	95	97	645	31	138	92	20	71	97	22	4	41	1	2	73
14052919_S11_L001	43	31	49	14	51	45	1	1	333	22	24	41	1	2	40
14057476_S17_L001	30	31	33	23	26	23	20	20	208	21	15	24	1	17	1
14058444_S18_L001	135	125	235	23	215	80	32	81	87	28	4	140	1	32	24
14058458_S19_L001	129	122	153	299	621	336	32	40	125	330	4	75	1	28	13
14061139_S21_L001	496	73	82	281	573	102	18	14	106	53	33	37	1	41	55
14061942_S22_L001	1	31	33	79	119	1	1	1	84	1	1	28	1	89	67
C140112_S24_L001	43	31	49	14	51	45	1	1	333	22	24	41	1	2	40
C140316_S27_L001	17	127	21	20	165	18	53	18	18	17	4	20	1	72	13
C140319_S30_L001	17	19	214	20	21	18	18	18	18	17	4	20	1	15	13
C140369_S17_L001	1	1	1	1	1	12	13	2	1	81	2	83	1	2	2
C140476_S18_L001	578	109	132	31	153	95	20	29	99	28	4	88	18	28	13
C140575-CAM_S23_L001	17	148	21	20	165	18	18	18	18	17	4	20	1	15	13
C140637_S22_L001	287	67	73	1	347	30	1	27	28	28	4	29	1	23	24
C140660_S23_L001	13	13	14	14	14	12	13	14	104	1	1	177	1	2	13
C140662-CAM_S59_L001	669	24	157	51	828	42	32	40	97	2	2	2	1	2	107
C140665-CAM_S95_L001	62	66	72	51	81	58	28	31	36	22	25	37	1	27	49
C140695-CAM_S84_L001	31	58	61	35	67	37	29	42	37	33	4	25	1	28	19
DTU2016_1950_PRJ1085_Campylobact	565	39	406	1	1	1	1	1	1	1	2	2	1	2	2
DTU2016_1951_PRJ1085_Campylobact	68	23	25	2	84	22	22	22	22	20	14	187	1	17	1
DTU2016_1959_PRJ1085_Campylobact	27	27	29	26	59	48	35	2	1	1	2	46	1	2	2
DTU2016_1960_PRJ1085_Campylobact	762	195	796	51	365	157	54	72	107	1	1	1	1	28	139
DTU2016_1961_PRJ1085_Campylobact	20	23	25	2	25	22	22	22	22	20	14	23	1	532	21
DTU2016_1962_PRJ1085_Campylobact	535	436	1488	670	1485	94	20	29	228	724	130	239	1	2	13
DTU2016_1963_PRJ1085_Campylobact	43	31	49	14	51	45	1	1	45	22	24	41	1	2	40
DTU2016_1964_PRJ1085_Campylobact	32	2	25	602	265	81	18	109	172	772	80	80	1	119	426
DTU2016_1965_PRJ1085_Campylobact	43	31	49	14	51	45	1	1	45	22	24	41	1	2	40

https://bitbucket.org/genomicepidemiology/cgmlstfinder/src/master/



https://bitbucket.org/genomicepidemiology/cgmlstfinder_db/src/master/

Campylobacter jejuni/coli (pubMLST)

Escheriachia coli, Salmonella, and Yersinia (EnteroBase)

SNP vs gene-by-gene approach

No nomenclature (SNP addresses has been indroduced by PHE)	cgST type
Comparability	Comparability
- Only by using the same reference genome	- higher as there is only one or a few cgMLST scheme
- Only by using the same SNP pipeline and parameters	
No limitation in term of reference genome	Limition by avaiable scheme
	- Public scheme have been developed and are maintained for many
	but not all pathogens
Restriced to regions of the genome present in all analyzed genomes	
useful information in the accessory genome is discarded	
Not suitable in scenarios where plasmid-mediated rather than clonal outbreak	

A.C. Church et al. Clinical Microbiology and Infection 24 (2018) 350-354

Isolate relatedness

Table 1
Examples of relatedness criteria for wg/cgMLST and SNP typing schemes of representative clinically relevant bacteria

Organism	Relatedness threshold[a]		References
	wg/cgMLST (allele) SNPs		
Acinetobacter baumannii	≤8	≤3	[25,26]
Brucella spp.	Epidemiologic validation in progress[b]		http://www.applied-maths.com/applications/wgmlst
Campylobacter coli, C. jejuni	≤14	≤15	[27,28]
Cronobacter spp.	Epidemiologic validation in progress[b]		http://www.applied-maths.com/applications/wgmlst
Clostridium difficile	Epidemiologic validation in progress[b]	≤4	[29], http://www.cgmlst.org/ncs, http://www.applied-maths.com/applications/wgmlst
Enterococcus faecium	≤20	≤16	[30]
Enterococcus raffinosus	Epidemiologic validation in progress[b]		http://www.applied-maths.com/applications/wgmlst
Escherichia coli	≤10	≤10	[31,32], https://enterobase.warwick.ac.uk/
Francisella tularensis	≤1	≤2	[33,34]
Klebsiella oxytoca	Epidemiologic validation in progress[b]		http://www.applied-maths.com/applications/wgmlst
Klebsiella pneumonia	≤10	≤18	[35,36]
Legionella pneumophila	≤4	≤15	[37]
Listeria monocytogenes	≤10	≤3	[38,39]
Mycobacterium abscessus		≤30	[40]
Mycobacterium tuberculosis	≤12	≤12	[41]
Neisseria gonorrhoeae	Epidemiologic validation in progress[b]	≤14	[42], http://www.applied-maths.com/applications/wgmlst
Neisseria meningitidis	Epidemiologic validation in progress[b]		http://www.cgmlst.org/ncs
Pseudomonas aeruginosa	≤14	≤37	[31,43]
Salmonella dublin	Epidemiologic validation in progress[b]	≤13	[44], https://enterobase.warwick.ac.uk/
Salmonella enterica	Epidemiologic validation in progress[b]	≤4	[45], http://www.cgmlst.org/ncs, http://www.applied-maths.com/applications/wgmlst, https://enterobase.warwick.ac.uk/
Salmonella typhimurium	Epidemiologic validation in progress[b]	≤2	[46], https://enterobase.warwick.ac.uk/
Staphylococcus aureus	≤24	≤15	[47,48]
Streptococcus suis		≤21	[49]
Vibrio parahaemolyticus	≤10		[50]
Yersinia spp.	0		[51]

A.C. Church et al. Clinical Microbiology and Infection 24 (2018) 350-354

Isolate relatedness

TABLE 1 | Maximum pairwise SNPs measured during investigations into foodborne illness outbreaks and contamination events.

Organism	Maximum SNP count (number)	Maximum SNP count (range)			Reference
		<21	21–100	>100	
E. coli	4	X			Underwood et al., 2013
E. coli	15	X			Eppinger et al., 2011
L. monocytogenes	9	X			Chen et al., 2017c
L. monocytogenes	12	X			Chen et al., 2017a
L. monocytogenes	18	X			Li et al., 2017
L. monocytogenes	20	X			Wang et al., 2015
L. monocytogenes	21		X		Nielsen et al., 2017
L. monocytogenes	28		X		Gilmour et al., 2010
L. monocytogenes	29		X		Chen et al., 2017b
L. monocytogenes	42		X		Chen et al., 2016
L. monocytogenes	67		X		Jackson et al., 2016
S. enterica	2	X			Wuyts et al., 2015
S. enterica	3	X			Allard et al., 2016
S. enterica	3	X			Taylor et al., 2015
S. enterica	6	X			Hoffmann et al., 2016
S. enterica	12	X			Octavia et al., 2015
S. enterica	30		X		Leekitcharoenphon et al., 2014

The maximum SNP counts for isolates that were traced back to the same source in the original study are presented. Whether the maximum SNP counts are less than 21 SNPs, 21 to 100 SNP, or greater than 100 SNPs is also indicated.

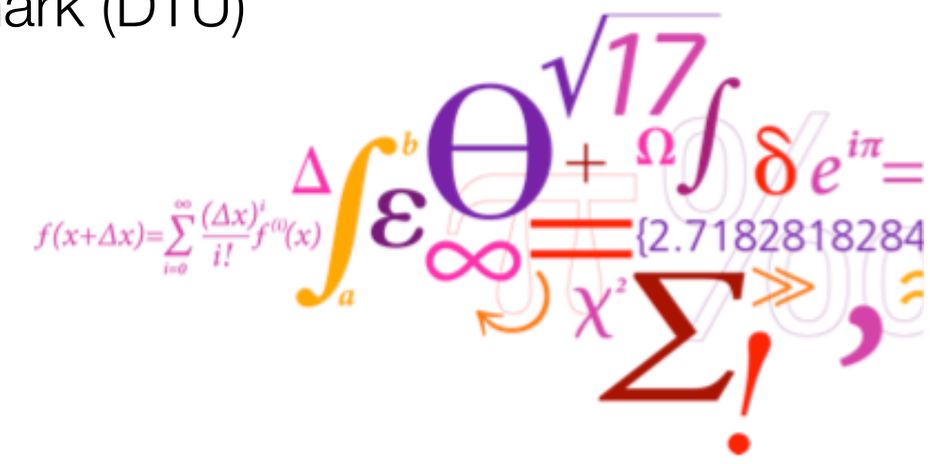
Pightling AW et al. Front. Microbiol. 9:1482 (2018)

Example of using cgMLST in a research study

GENCAMP project

Comparative genomics of quinolone-resistant *Campylobacter jejuni* of poultry origin from major poultry producing European countries

"Shinny" Pimlapas Leekitcharoenphon
Research Group for Genomic Epidemiology
DTU Food
Technical University of Denmark (DTU)



Leekitcharoenphon P, et al., 2018. EFSA Supporting Publications. https://doi.org/10.2903/sp.efsa.2018.EN-1398

Campylobacter

- Campylobacteriosis is a predominant bacterial cause of gastroenteritis worldwide

- In UK and the USA, the majority (90%) of human disease is caused by *Campylobacter jejuni*

- *Campylobacter jejuni* is in many countries the most important foodborne bacterial species

- They are found at high prevalence in commercial broiler chickens

- Contaminated poultry meat is a predominant source of human *Campylobacter* infection

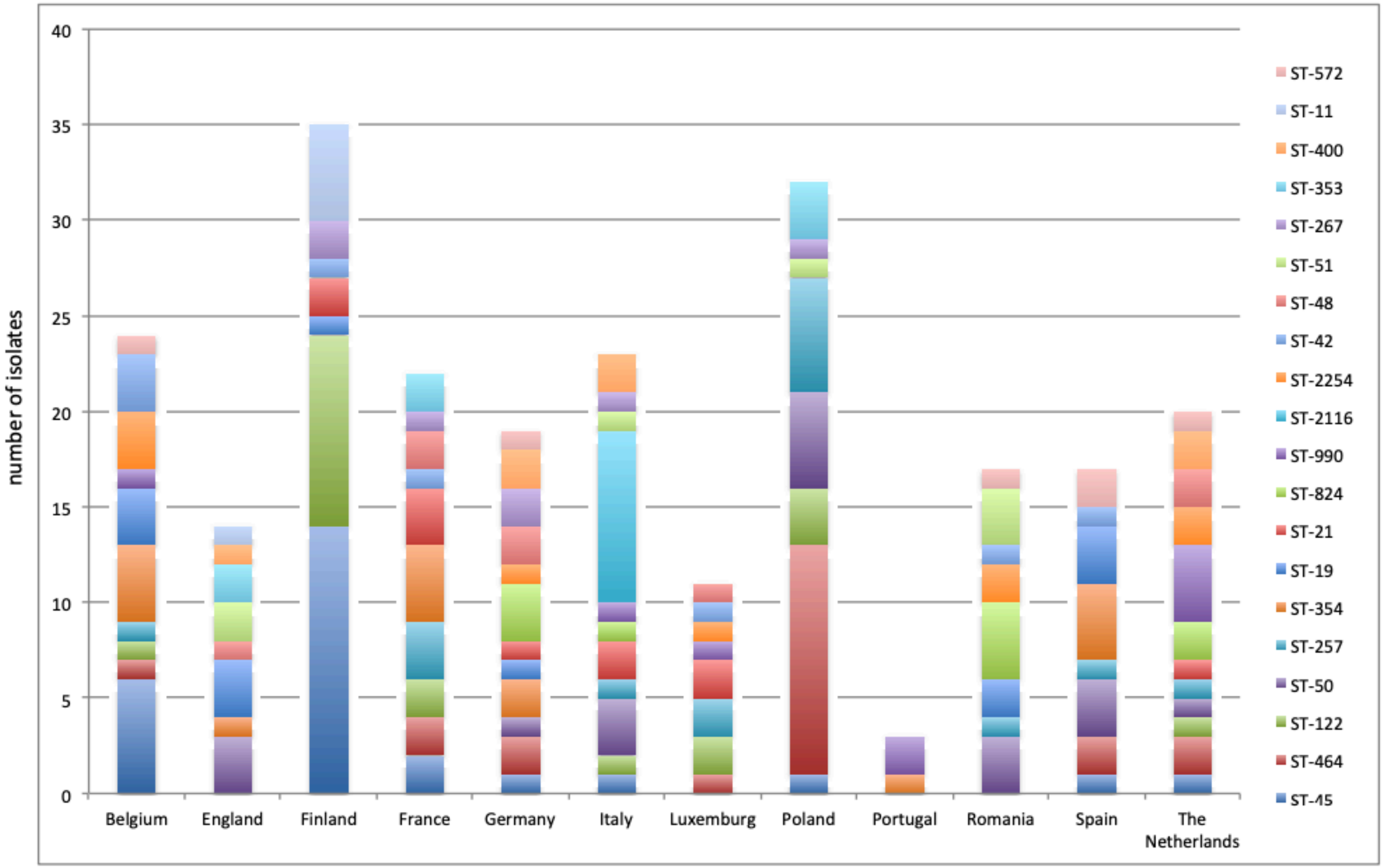
GenCamp

- To examine the genomic diversity of quinolone resistant and susceptible *C. jejuni* across the poultry producing European countries

- The emergence of quinolone resistance among *C. jejuni* is related to transmission through countries OR selected though quinolone use in the individual countries

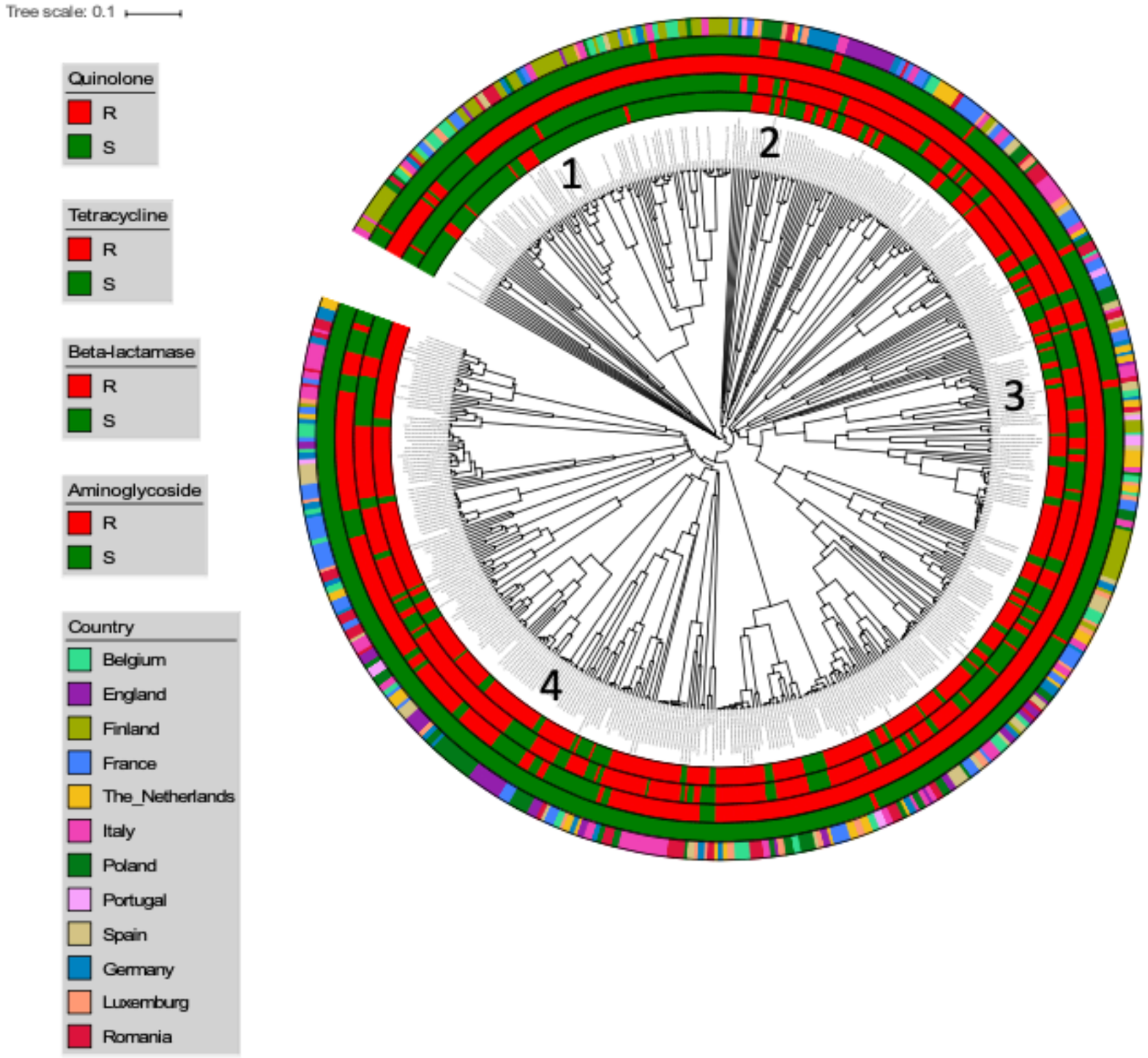
GenCamp

- A total of 502 *C. jejuni* isolates from poultry in 12 European countries
- 307 isolates contained *gyr*A mutation conferring quinolone resistance
- 144 different ST-types were observed among isolates
- cgMLST tree, wgMLST tree, SNP tree and ND tree



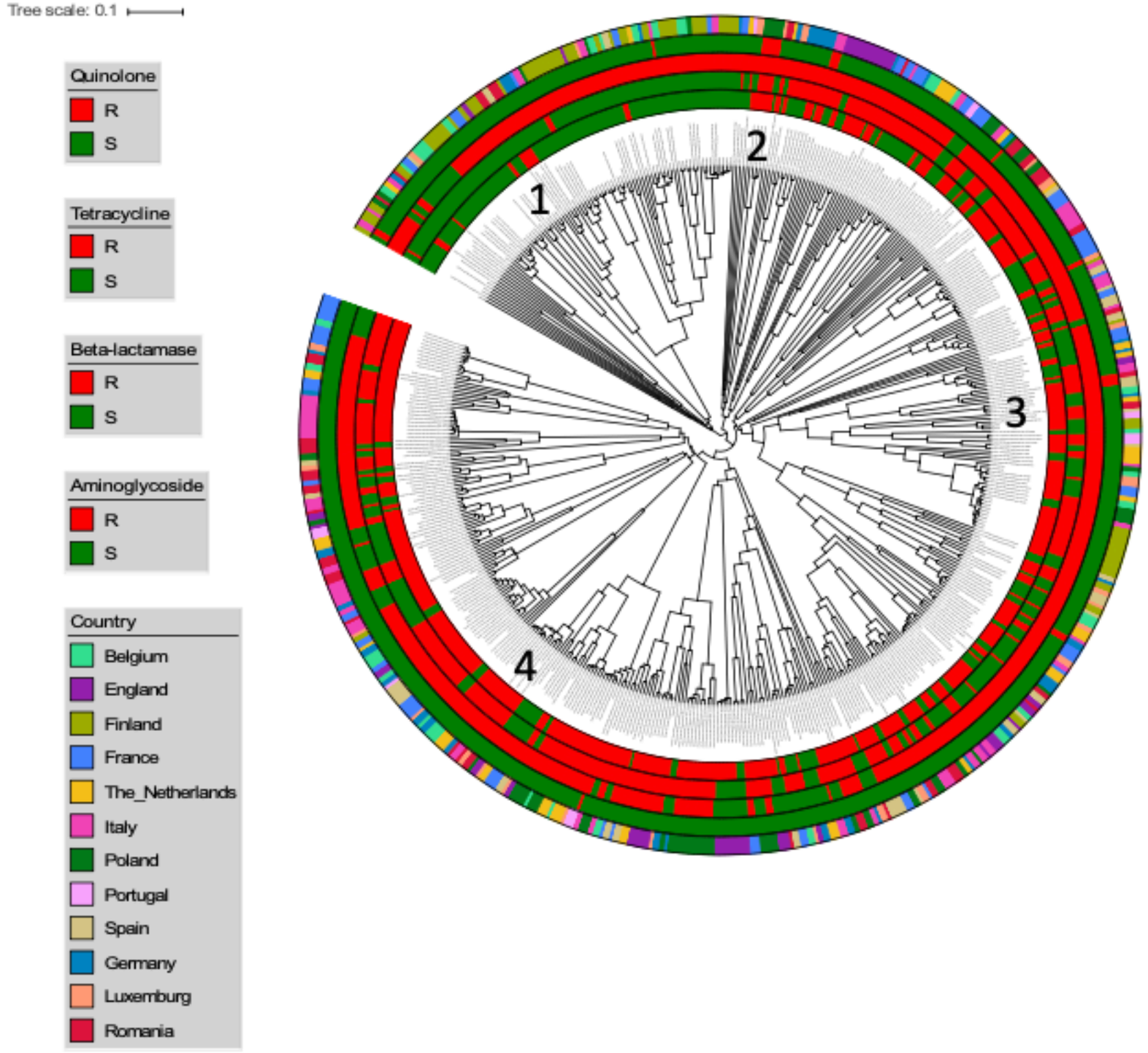
Y-axis is the number of strains.

Figure 1: Distribution of the top 20 ST-types (conventional seven multilocus sequence types)



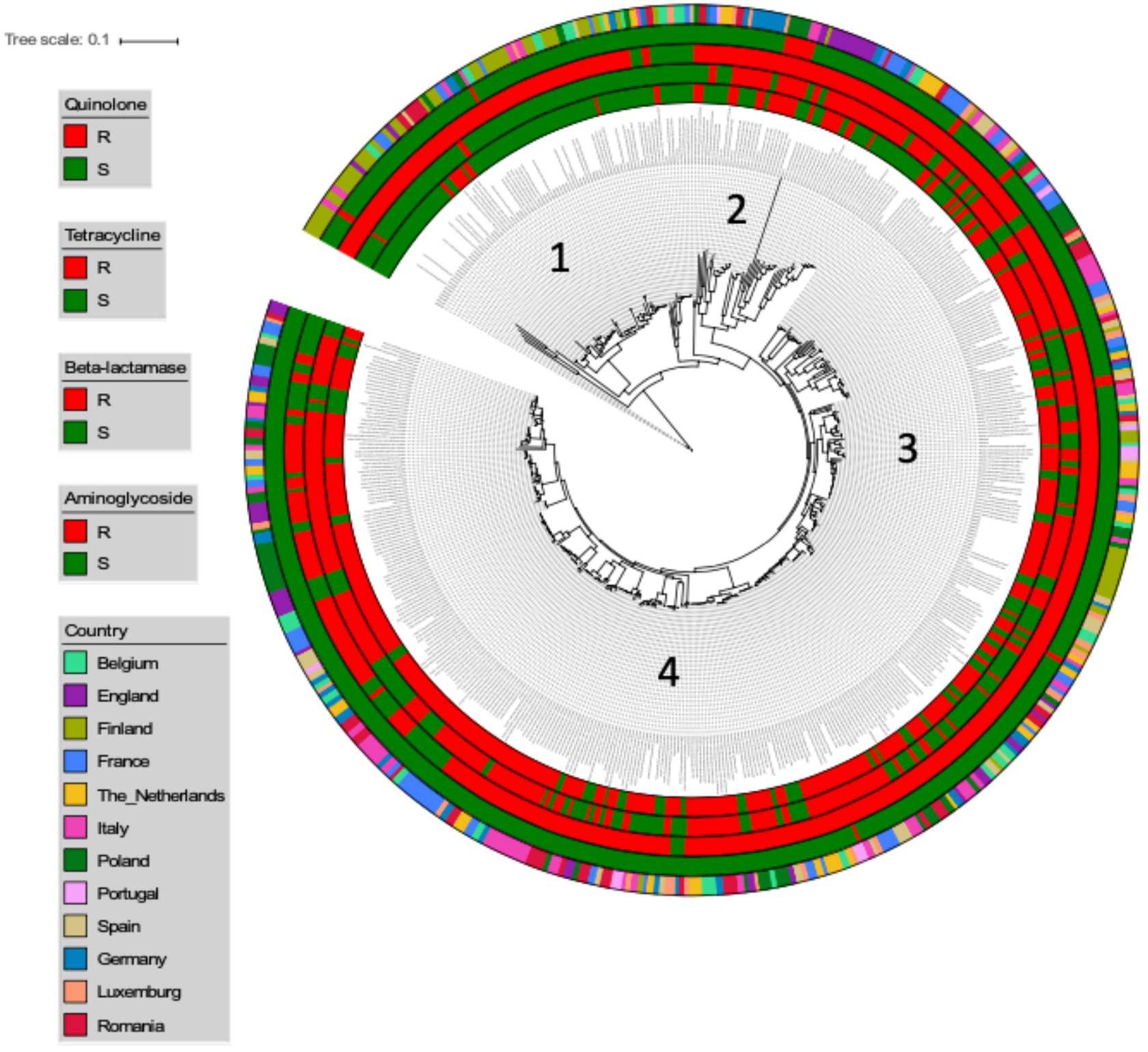
The order of the circle lanes from inner to outer was quinolone (R: resistance caused by *gyr*A mutations; S: no *gyr*A mutation detected), tetracycline (R: resistance caused by *tet*(O); S: no *tet*(O) gene), beta-lactamase (R: resistance caused by *bla*OXA gene; S: no *bla*OXA gene), aminoglycoside (R: resistance caused by *aph*(3')-III, *aad*E; S: no *aph*(3')-III, *aad*E genes) and country.

Figure 2: Core genome MLST tree of the 502 isolates



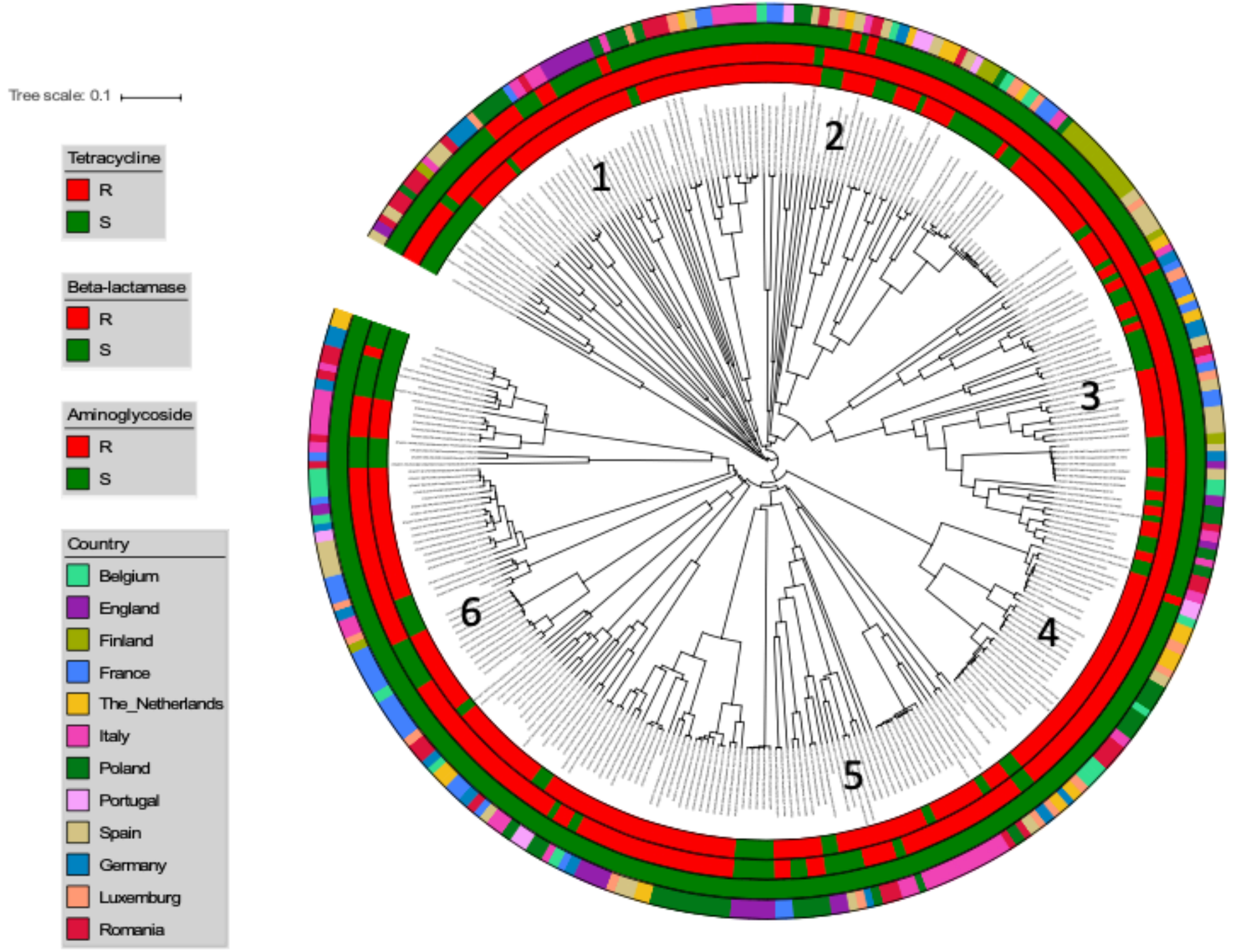
The order of the circle lanes from inner to outer was quinolone (R: resistance caused by gyrA mutations; S: no gyrA mutation detected), tetracycline (R: resistance caused by *tet*(O); S: no *tet*(O) gene), beta-lactamase (R: resistance caused by *bla*OXA gene; S: no *bla*OXA gene), aminoglycoside (R: resistance caused by *aph*(3')-III, *aad*E; S: no *aph*(3')-III, *aad*E genes) and country.

Figure 3: Whole genome MLST tree of the 502 isolates



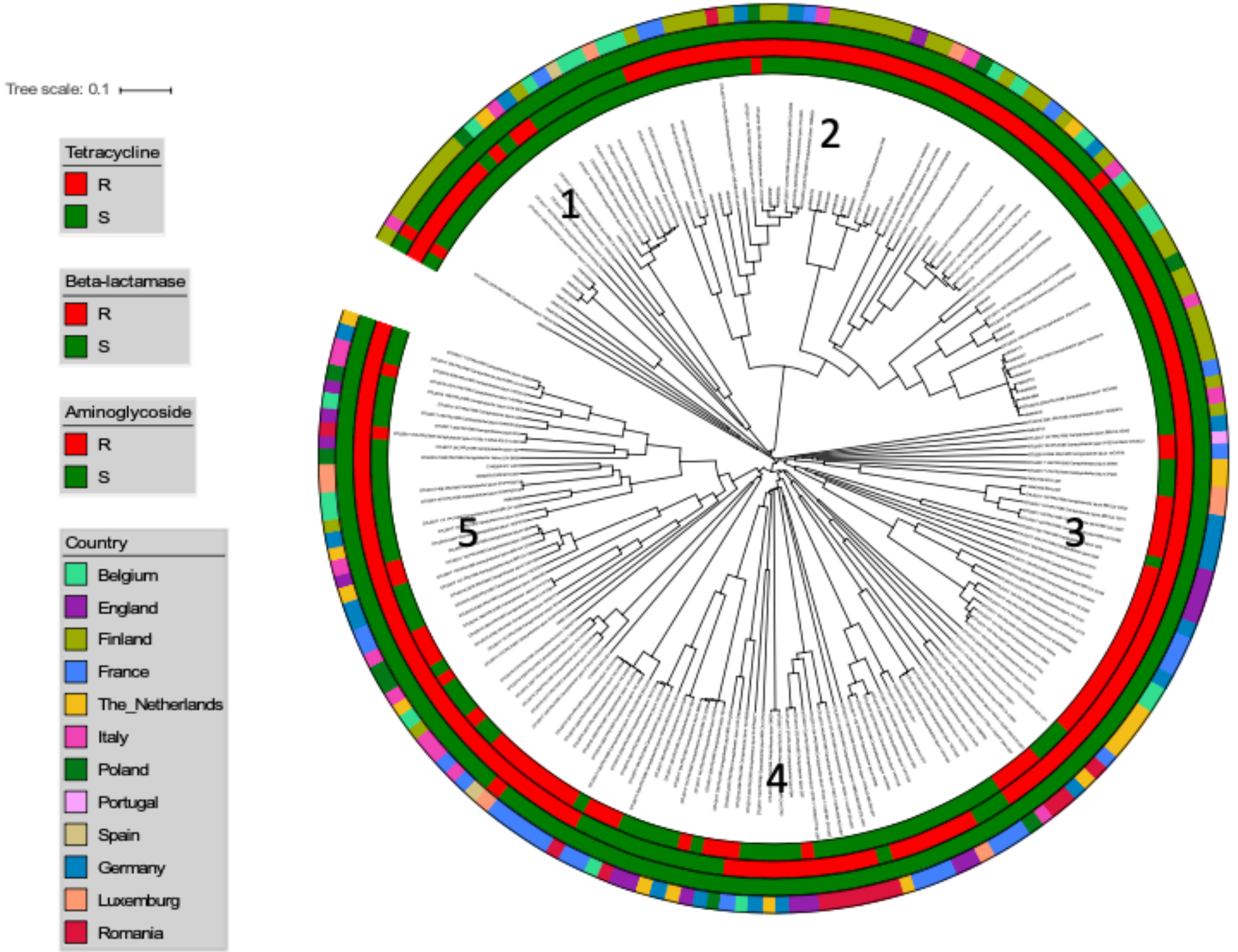
The order of the circle lanes from inner to outer was quinolone (R: resistance caused by *gyr*A mutations; S: no *gyr*A mutation detected), tetracycline (R: resistance caused by *tet*(O); S: no *tet*(O) gene), beta-lactamase (R: resistance caused by *bla*OXA gene; S: no *bla*OXA gene), aminoglycoside (R: resistance caused by *aph*(3')-III, *aad*E; S: no *aph*(3')-III, *aad*E genes) and country.

Figure 4: SNP tree of the 502 isolates



The order of the circle lanes from inner to outer was tetracycline (R: resistance caused by *tet*(O); S: no *tet*(O) gene), beta-lactamase (R: resistance caused by *bla*OXA gene; S: no *bla*OXA gene), aminoglycoside (R: resistance caused by *aph*(3')-III, *aad*E; S: no *aph*(3')-III, *aad*E genes) and country.

Figure 9: cgMLST tree of the quinolone resistant isolates



The order of the circle lanes from inner to outer was tetracycline (R: resistance caused by *tet*(O); S: no *tet*(O) gene), beta-lactamase (R: resistance caused by *bla*OXA gene; S: no *bla*OXA gene), aminoglycoside (R: resistance caused by *aph*(3')-III, *aad*E; S = no *aph*(3')-III, *aad*E genes) and country.

Figure 10: cgMLST tree of the quinolone susceptible isolates

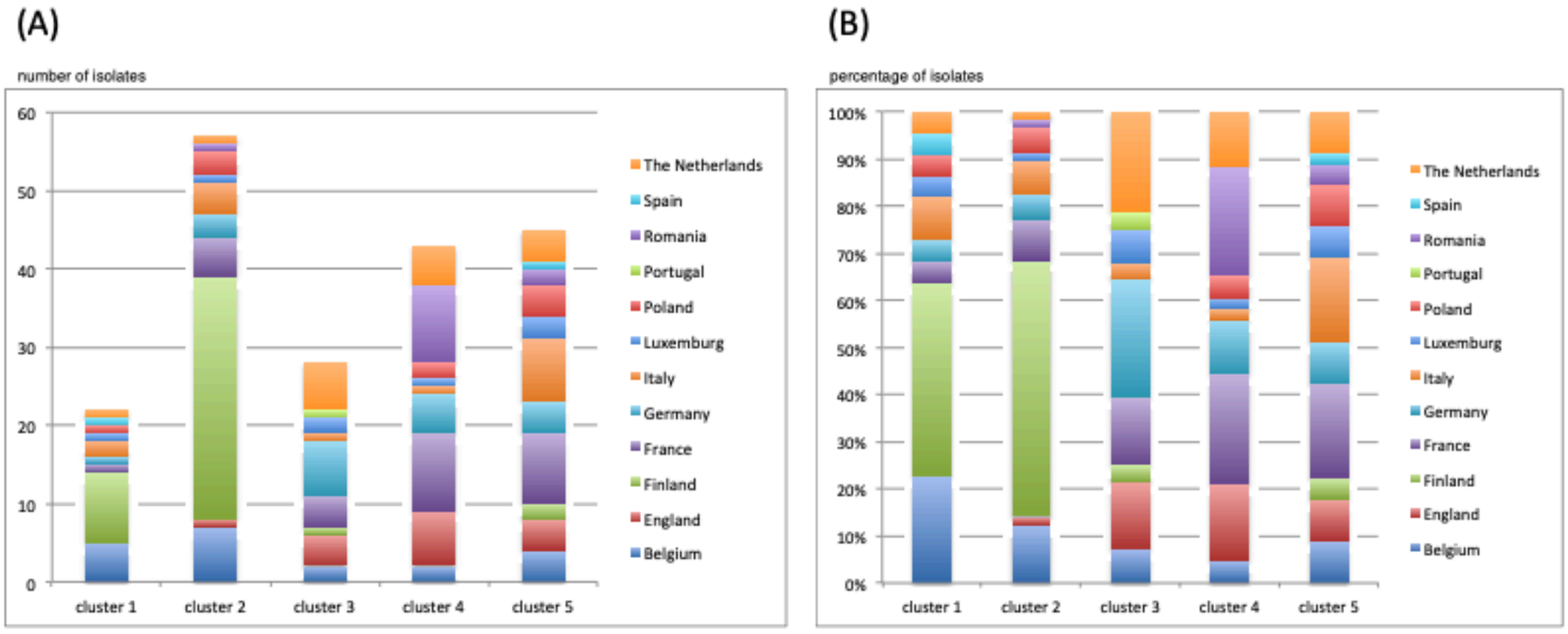


Figure 11: Distribution of FQ-S isolates in the different clusters (A and B) and in different countries (C and D) defined by the cgMLST tree

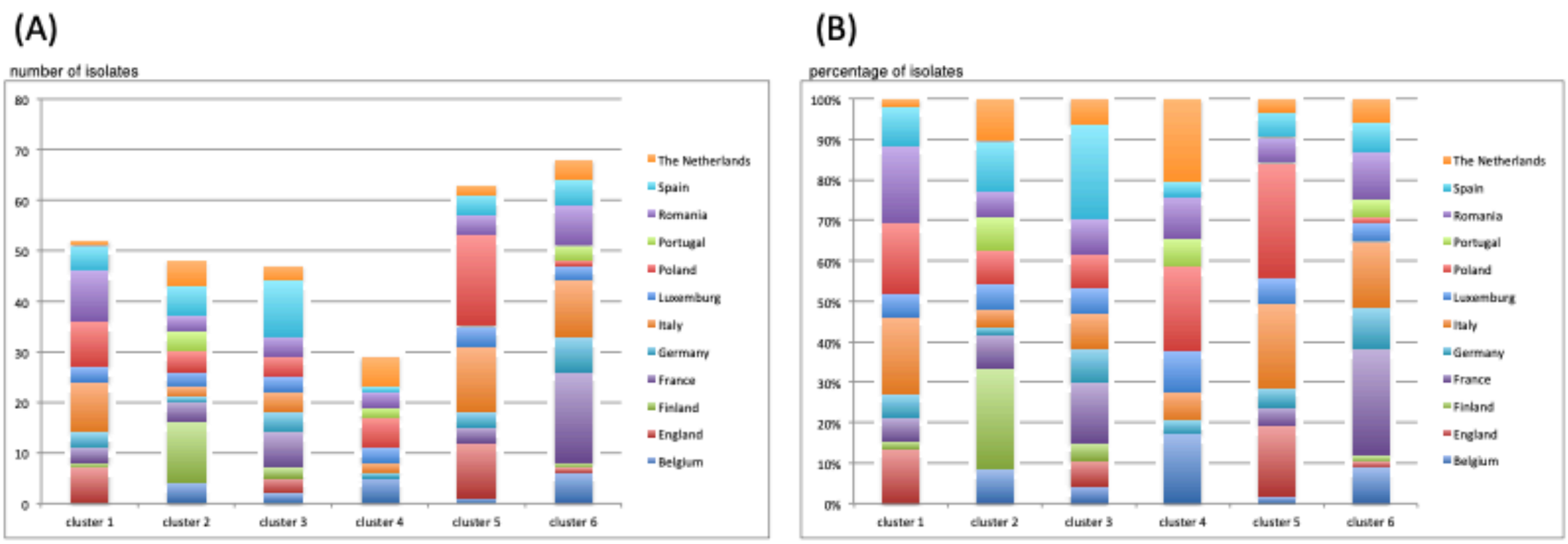


Figure 12: Distribution of the FQ-R isolates in the different clusters (A and B) and in different countries (C and D) defined by the cgMLST tree

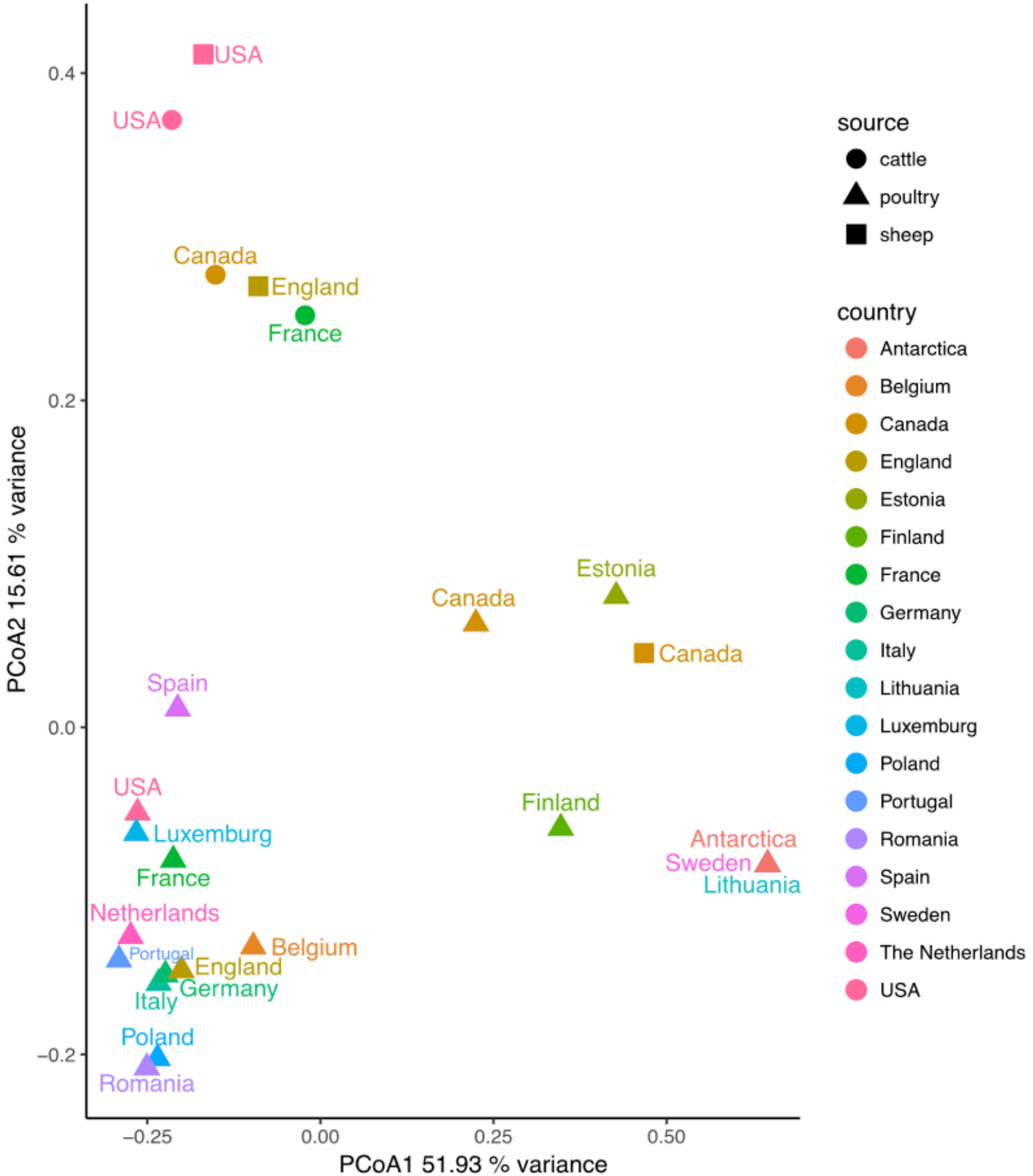


Figure 13: Principal coordinate analysis (PCoA) of allele frequency at country and host levels

Assembly vs Raw read

- What would you use between assembled genome (FASTA) or raw reads (FASTQ) to build SNP tree and cgMLST tree ?

- Can you use assembled genomes for building SNP tree and cgMLST tree ?

Exercise
Outbreak blusters

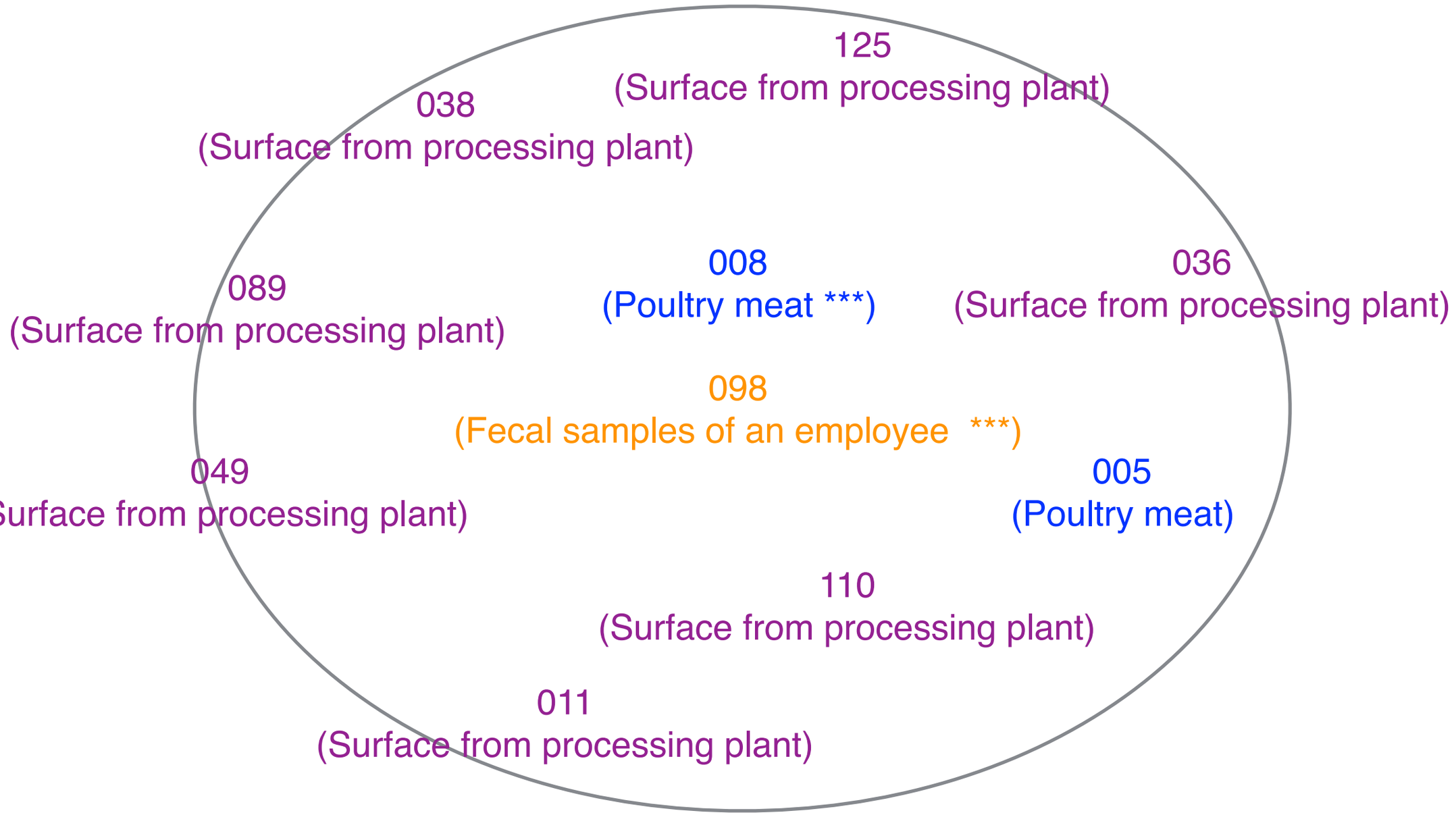
Objectives

- Perform cgMLST analysis
- Assess the relatedness of isolates based on phylogenetic trees (cgMLST)
- Hypothesize possible epidemiological scenarios of transmission of AMR isolates based on the trees

- **Case study:** The two colistin-resistant mcr-1-positive *Escherichia coli* ST131 strains were isolated from a poultry meat sample (strain ID 008) and a fecal sample of an employee at the poultry meat processing plant (strain ID 098). During a three-month investigation to assess the occurrence of possible zoonotic agents in a meat processing plant, further Escherichia coli ST131 were isolated from poultry meat and from surfaces of the processing plant (see table below)

Source	Strain ID
Poultry meat	005, 008
Surfaces from processing plant	011, 036, 038, 049, 089, 110, 125
Fecal samples of an employee	098

Case story of pathogenic *E.coli* ST131



- genetic characterisation of the strains

- identify possible transmission pathways

```
/home/projects/course_23262/course/week06/cgmlst/raw_trimmed_data
```

***Escherichia coli* ST131** includes some of the most prevalent extra-intestinal pathogenic *E. coli* clones globally, causing a wide range of infections (e.g., septicemia and urinary tract infections). These *E. coli* clones can colonize the intestinal tract of humans and food-producing animals, and there is mounting evidence that at least a subset of *E. coli* ST131 of animal origin might reach the human intestinal tract via consumption of meat and cause extra-intestinal infections.

For further reading please see Liu et al., mBio 2018 "*Escherichia coli* ST131-*H*22 as a foodborne uropathogen".

Questions

- 1) Based on the metadata provided and on the phylogenetic trees, what is the most likely epidemiological scenario?

- 2) Do the phylogenetic trees contain enough information for you to conclude anything about the epidemiology of the mcr-1 gene? Please motivate you answer.

How to build cgMLST tree

- cgMLST allele matrix using cgMLST Finder
- cgMLST tree using R
- Tree visualisation using iTol

cgMLST Finder

https://bitbucket.org/genomicepidemiology/cgmlstfinder/src/master/

Usage

The program can be invoked with the -h option to get help and more information of the service.

```
usage: cgMLST.py [-h] -s SPECIES -db DB_DIR [-o OUTPUT_DIR] [-t TMP_DIR]
                 [-k KMA_PATH] [-n NJ_PATH] [-mem]
                 FASTQ [FASTQ ...]


positional arguments:
  FASTQ                 FASTQ files to do cgMLST on.

optional arguments:
  -h, --help            show this help message and exit
  -s SPECIES, --species SPECIES
                        Species. Must match the name of a species in the
                        database
  -db DB_DIR, --databases DB_DIR
                        Directory containing the databases and gene lists for
                        each species.
  -o OUTPUT_DIR, --outdir OUTPUT_DIR
                        Output directory.
  -t TMP_DIR, --tmp_dir TMP_DIR
                        Temporary directory for storage of the results from
                        the external software.
  -k KMA_PATH, --kmapath KMA_PATH
                        Path to executable kma program.
  -n NJ_PATH, --nj_path NJ_PATH
                        Path to executable neighbor joining program.
  -mem, --shared_memory
                        Use shared memory to load database.
```

Example of command to run cgMLSTFinder:

```
python3 cgMLST.py /path/to/isolate.fq.gz -s ecoli -o /path/to/outdir -db /path/to/cgmlstfinder_db/
-k /usr/local/bin/kma
```

cgMLST_table.txt

Genome	AEJV01_03887	C_RS24035	C_RS24040	EAKF1_RS0794	ECs0267	ECs4266	ECs4518
110_R1	2	143	1	11	43	76	71
125_R1	2	143	1	11	43	76	71
038_R1	2	143	1	11	43	76	71
049_R1	2	143	1	10	43	76	71
005_R1	2	143	1	11	43	76	71
008_R1	2	143	1	44	43	76	71
089_R1	2	143	1	11	43	76	71
098_R1	2	143	1	44	43	76	71
011_R1	2	143	1	11	43	76	71
036_R1	2	143	1	11	43	76	71

How to build cgMLST tree

- cgMLST allele matrix using cgMLST Finder

- cgMLST tree using R

  - Link to Rmarkdown  Rmarkdown_cgMLST_tree.rmd

cgMLST allele matrix at DTU Learn (ecoli_results.txt)
cgMLST ST type (cgST) at DTU Learn (ecoli_summary.txt)

```
---
title: 'cgMLST tree'
author: "Shinny Pimlapas Leekitcharoenphon"
date: "September 21, 2018"
output:
  html_document:
    theme: sandstone
    code_folding: hide
---


```{r init, message=FALSE}
knitr::opts_chunk$set(cache = TRUE, autodep = TRUE, warning=FALSE, message=FALSE)
#install.packages("cluster")
library(cluster)
#install.packages("ape")
library(ape)
#install.packages("reshape")
library(reshape)
```
cgMLST tree (distance not included) {#E2A}
```{r cgMLST tree, eval=TRUE}
data <- read.table("cgMLST_table.txt", sep = "\t", row.names=1, colClasses = "factor", header = T)
cgMLST_tree <- as.phylo(hclust(daisy(data, metric="gower")))
write.tree(phy=cgMLST_tree, file="cgMLST_tree.newick")
plot(hclust(daisy(data, metric="gower")))

```


cgMLST dissimilarity table {#E2A}
```{r dissimilarity tree, eval=TRUE}

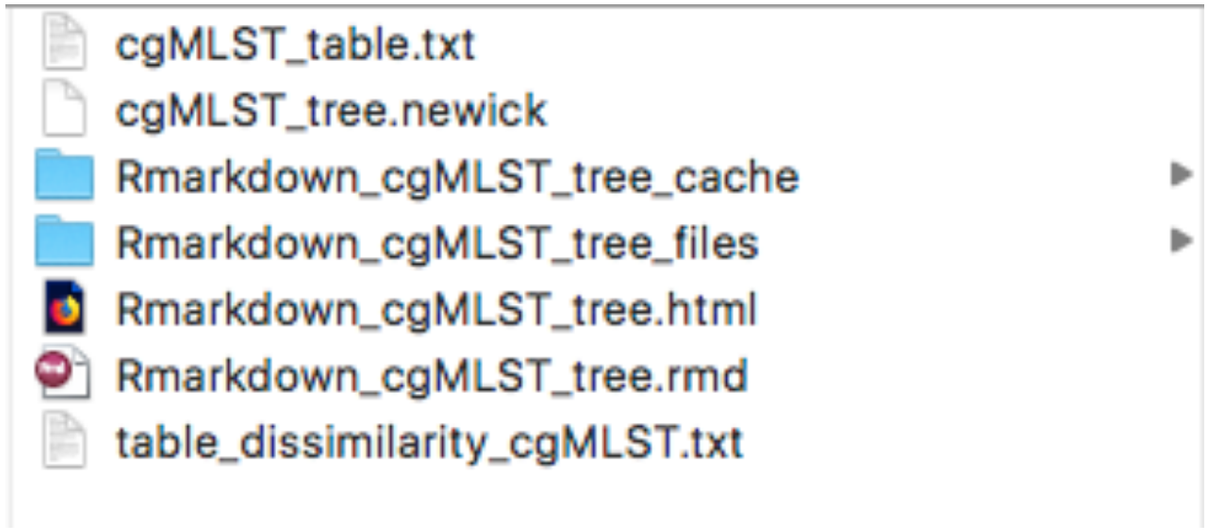
m <- as.matrix(daisy(data, metric="gower"))
m2 <- melt(m)[melt(upper.tri(m))$value,]
names(m2) <- c("c1", "c2", "distance")
m
m2
write.table(m, 'table_dissimilarity_cgMLST.txt', sep='\t')
```
```

Install following packages in R before running the script

```
install.packages("cluster")

install.packages("ape")

install.packages("reshape")
```




```yaml
---
title: 'cgMLST tree'
author: "Shinny Pimlapas Leekitcharoenphon"
date: "September 21, 2018"
output:
  html_document:
    theme: sandstone
    code_folding: hide
---
```

~/Meetings/175_EURL_September2018/exercise3/cgMLST/c

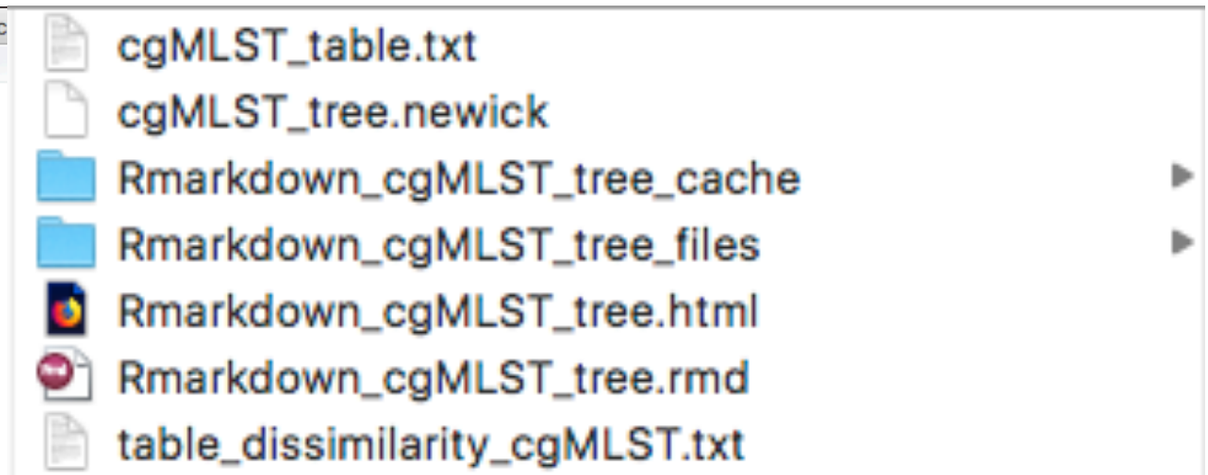
Rmarkdown_cgMLST_tree.html | Open in Browser | Find

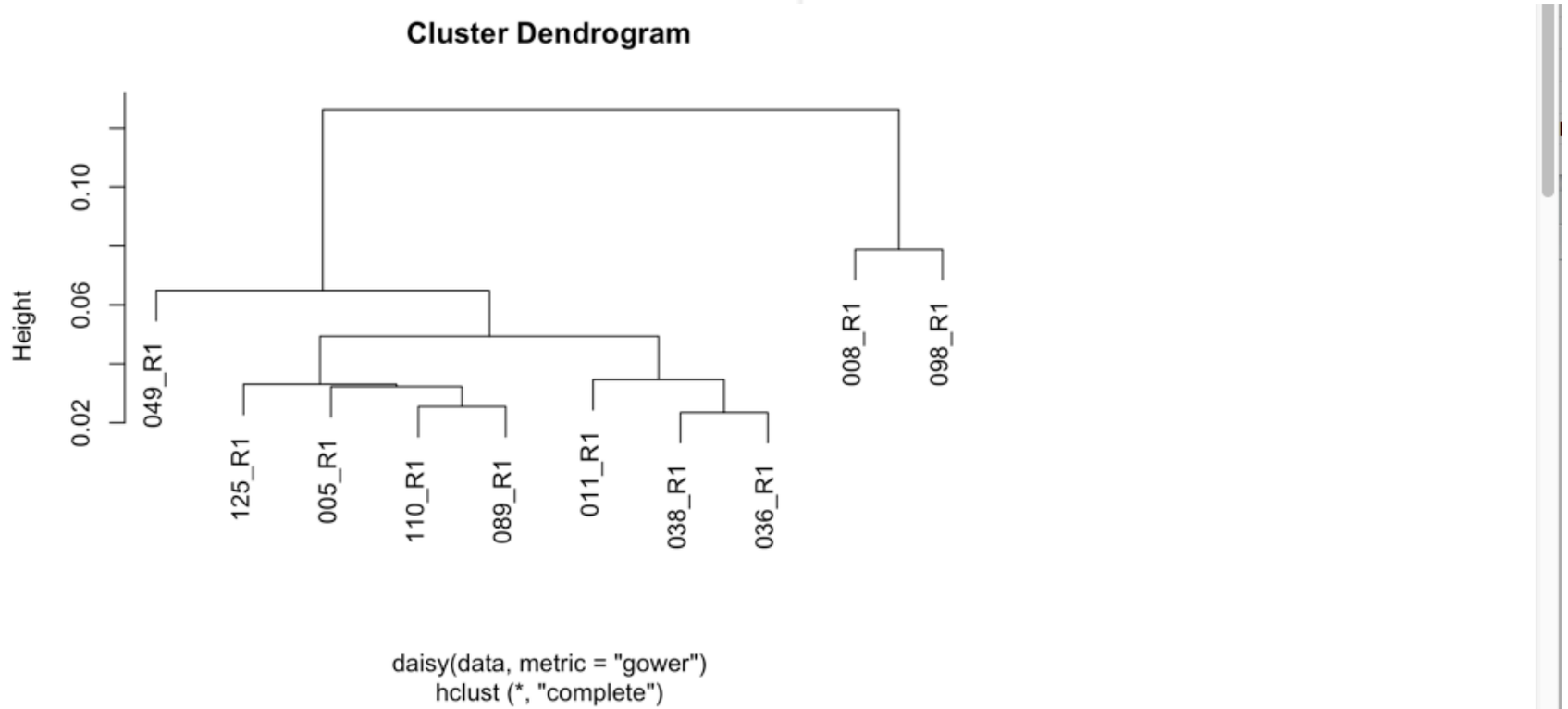
cgMLST tree

Shinny Pimlapas Leekitcharoenphon

September 21, 2018

cgMLST tree (distance not included)





cgMLST dissimilarity table

CODE

```
##            110_R1     125_R1     038_R1     049_R1     005_R1     008_R1
## 110_R1 0.00000000 0.02666136 0.03461998 0.04974135 0.02785515 0.11221647
## 125_R1 0.02666136 0.00000000 0.04058894 0.05531238 0.03263032 0.11778750
## 038_R1 0.03461998 0.04058894 0.00000000 0.05531238 0.03302825 0.12017509
## 049_R1 0.04974135 0.05531238 0.05531238 0.00000000 0.04974135 0.12614405
```

table_dissimilarity_cgMLST.txt

110_R1	125_R1	038_R1	049_R1	005_R1	008_R1	089_R1	098_R1	011_R1	036_R1	
110_R1	0	0,026661361	0,034619976	0,049741345	0,027855153	0,112216474	0,025467569	0,097095105	0,042180661	0,030242738
125_R1	0,026661361	0	0,040588938	0,055312376	0,032630322	0,117787505	0,033028253	0,102666136	0,047751691	0,035415838
038_R1	0,034619976	0,040588938	0	0,055312376	0,033028253	0,12017509	0,037405491	0,105053721	0,034222045	0,023477915
049_R1	0,049741345	0,055312376	0,055312376	0	0,049741345	0,126144051	0,05252686	0,111022682	0,064862714	0,050935137
005_R1	0,027855153	0,032630322	0,033028253	0,049741345	0	0,11539992	0,032232392	0,101074413	0,041384799	0,029844807
008_R1	0,112216474	0,117787505	0,12017509	0,126144051	0,11539992	0	0,116195782	0,07879029	0,124552328	0,114206128
089_R1	0,025467569	0,033028253	0,037405491	0,05252686	0,032232392	0,116195782	0	0,101472344	0,049343414	0,034222045
098_R1	0,097095105	0,102666136	0,105053721	0,111022682	0,101074413	0,07879029	0,101472344	0	0,111022682	0,101074413
011_R1	0,042180661	0,047751691	0,034222045	0,064862714	0,041384799	0,124552328	0,049343414	0,111022682	0	0,034619976
036_R1	0,030242738	0,035415838	0,023477915	0,050935137	0,029844807	0,114206128	0,034222045	0,101074413	0,034619976	0

Allele difference = cgMLST dissimilarity x total alleles

How to build cgMLST tree

- cgMLST allele matrix using cgMLST Finder
- cgMLST tree using R
- Tree visualisation using iTol

Tree file in newick format at DTU Learn (cgMLST_tree.newick)

cgMLST dissimilarity matrix at DTU Learn (table_dissimilarity_cgMLST.txt)