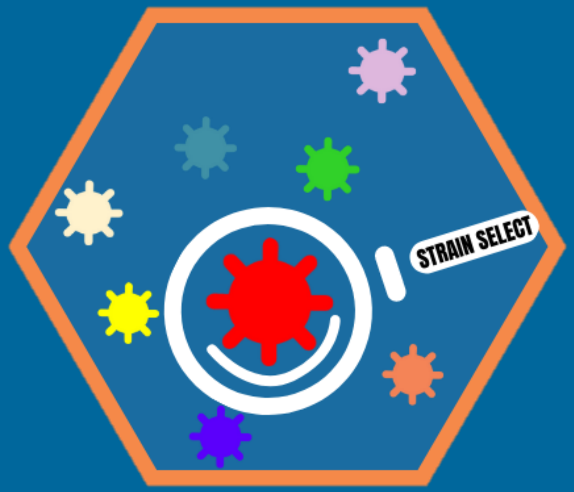
Strain select Package: a R-tool for strain selection



L. Guillier

Laurent.guillier@anses.fr

OHEJP WP4 Dissemination Webinar on Surveillance and Risk Assessment
28 March 2023

This presentation is part of the European Joint Programme One Health EJP. This project has received funding from the European Union’s Horizon 2020 research and innovation programme under Grant Agreement No 773830.

The Strain select R package

Impact

Aim: Propose a formal and reproducible method to select strains in a collection based on their metadata

Areas on applications:

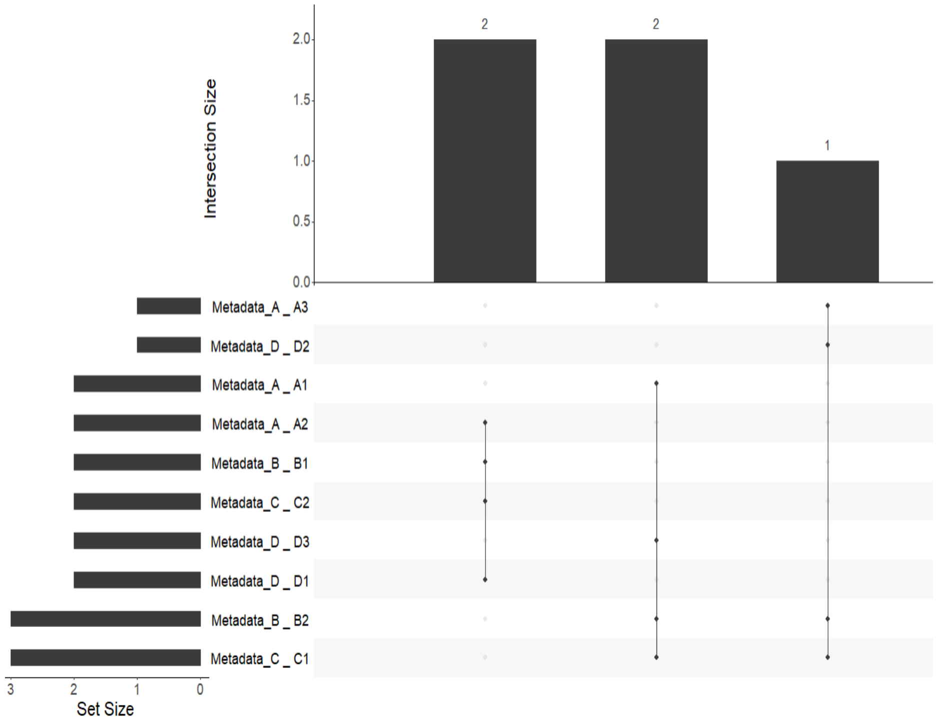
- Any pathogens
- Many contexts in One Health studies
  - Choosing strains to be sequenced for epi. investigations
  - Establishing a set of strains for phenotyping
  - Selecting strains representative of consumer exposure or risk
  - Selecting strains based on their virulence/antimicrobial profiles
  - …

Added values

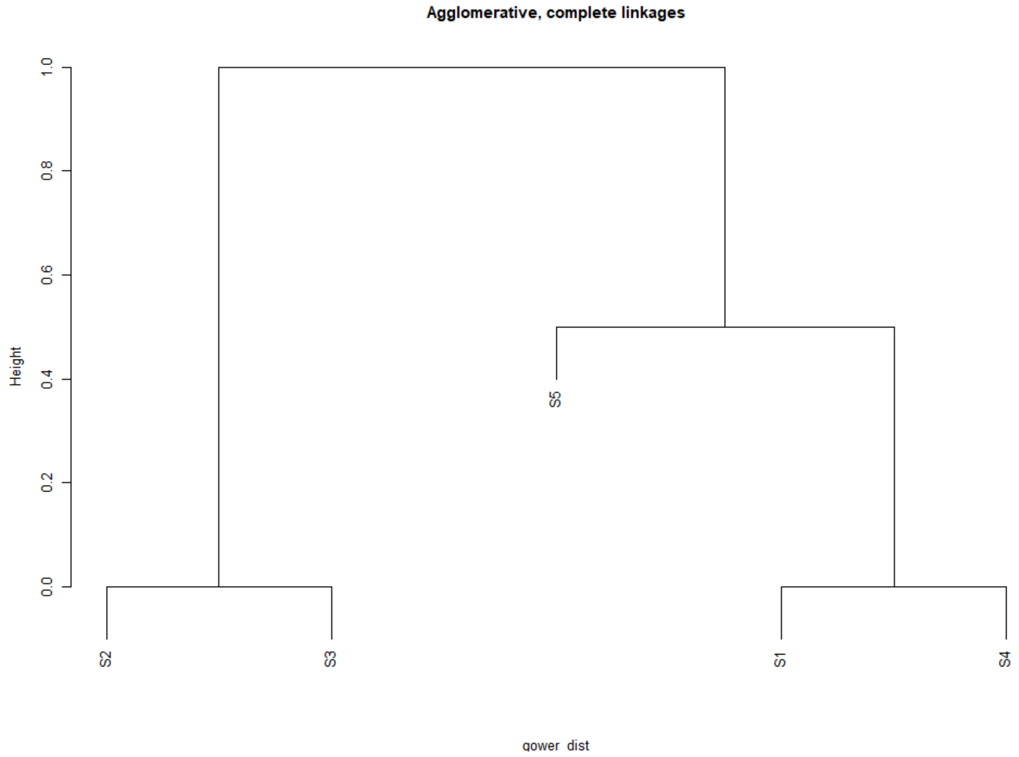
- Reference laboratories (NL- EU-)
- Research teams working on population structure, source attribution

Strain select R package: Two methods for selecting strains

❶ Intersection-based method



❷ Clustering method (Gower)



Strain select R package: Two methods for selecting strains

Same structure of data frame for the two methods

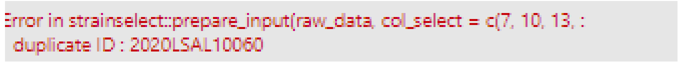
prepare_input(filename, metadata)

Mystrains.xlsx

Selection of metadata (column numbers)

Outputs:

- quality checks (e.g. duplicate strain names)
- data input for selection methods



```
Error in strainselect::prepare_input(raw_data, col_select = c(7, 10, 13, :
  duplicate ID : 2020LSAL10060
```

Strain ID	Metadata_A	Metadata_B	Metadata_C	Metadata_D
S1	A1	B2	C1	D2
S2	A2	B1	C2	D1
S3	A2	B1	C2	D1
S4	A1	B2	C1	D2
S5	A2	B2	C2	D2

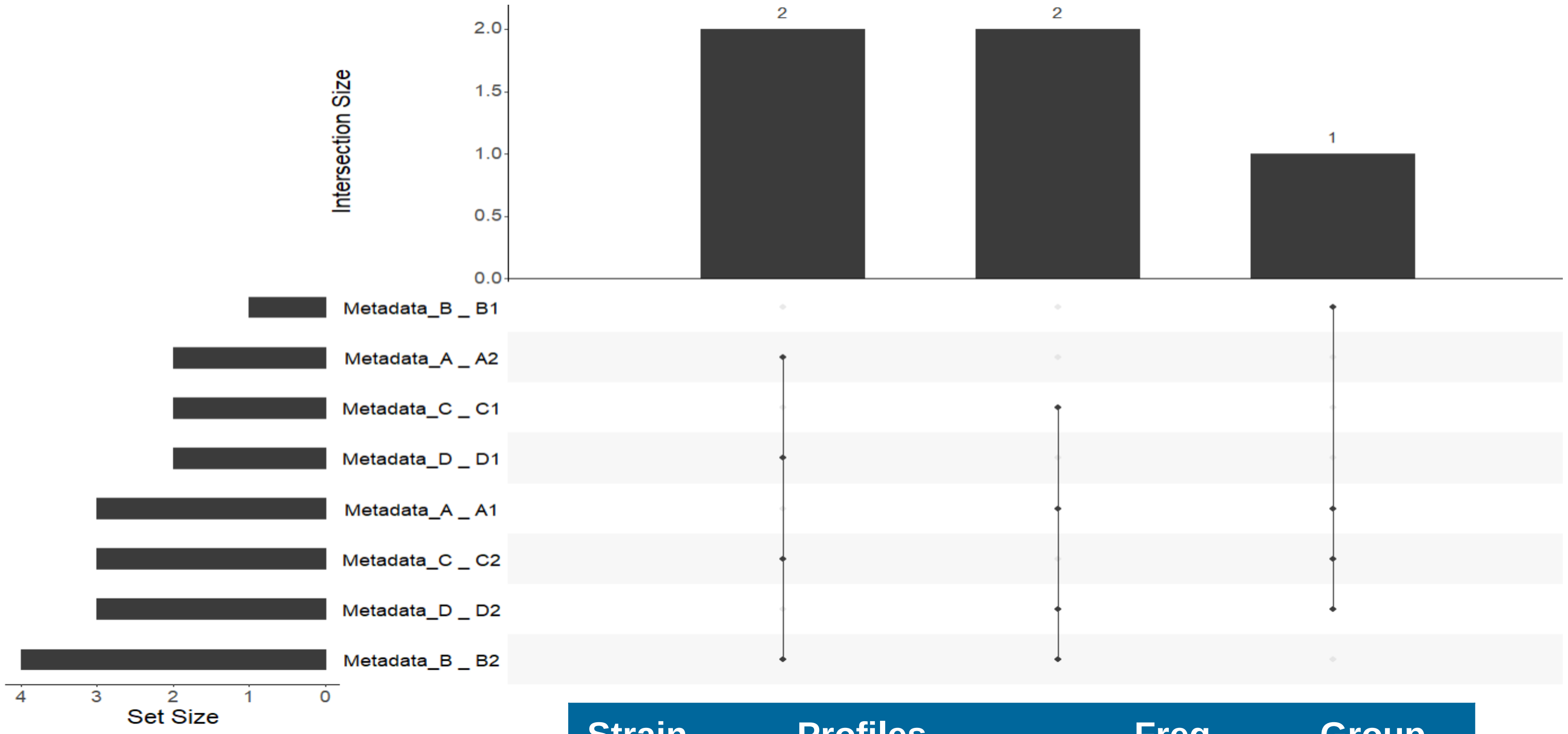
Strain select R package: ❶ Intersection-based method

Strain ID	Metadata_A	Metadata_B	Metadata_C	Metadata_D
S1	A1	B2	C1	D2
S2	A2	B1	C2	D1
S3	A2	B1	C2	D1
S4	A1	B2	C1	D2
S5	A2	B2	C2	D2

prepare_upset()

Strain ID	Metadata_A_A1	Metadata_A_A2	Metadata_B_B1	...	Metadata_D_D2
S1	1	0	0	...	1
S2	0	1	1	...	0
S3	0	1	1	...	0
S4	1	0	0	...	1
S5	0	1	0	...	1

Strain select R package: ❶ Intersection-based method



Strain	Profiles	Freq	Group
S1	10011001	2	1
S2	01100110	2	2
S3	01100110	2	2
S4	10011001	2	1
S5	01010101	1	3

define_profiles() and select_profiles()

→ Random selection in groups

Group	Selected strain
1	S4
2	S3
3	S5

Strain select R package: ❷ Clustering method (Gower)

Strain	Metadata_A	Metadata_B	Metadata_C	Metadata_D
S1	A1	B2	C1	D2
S2	A2	B1	C2	D1
S3	A2	B1	C2	D1
S4	A1	B2	C1	D2
S5	A2	B2	C2	D2

assess_gower()

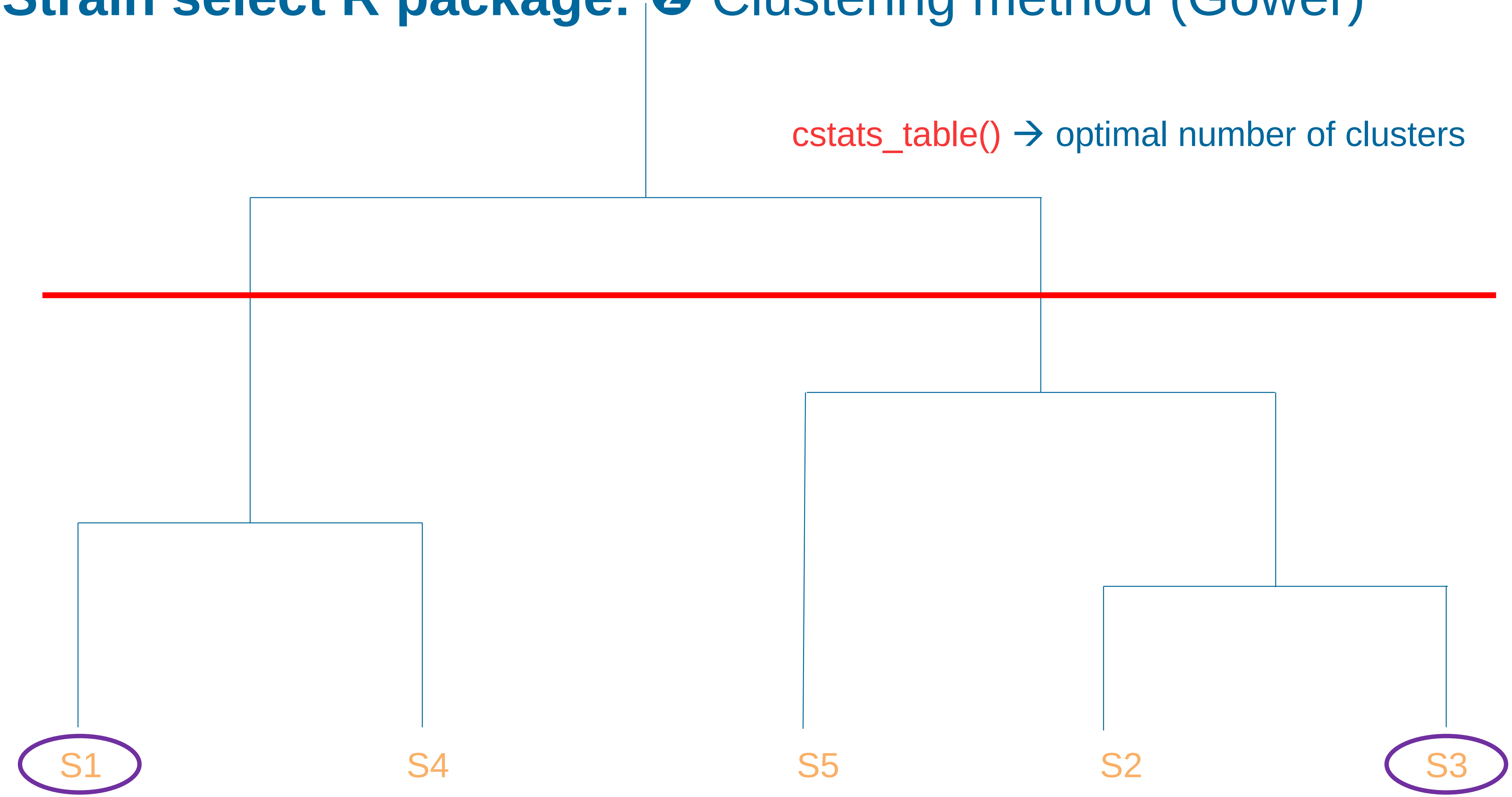
$$D_{Gower}(x_i, x_j) = 1 - S_{Gower}(x_i, x_j)$$

$$S_{Gower}(x_i, x_j) = \frac{\sum_{k=1}^{p} s_{ijk}}{p}$$

	S1	S2	S3	S4	S5
S1	0	4	4	0	2
S2		0	0	4	2
S3			0	4	3
S4				0	2
S5					0

Strain select R package: ❷ Clustering method (Gower)

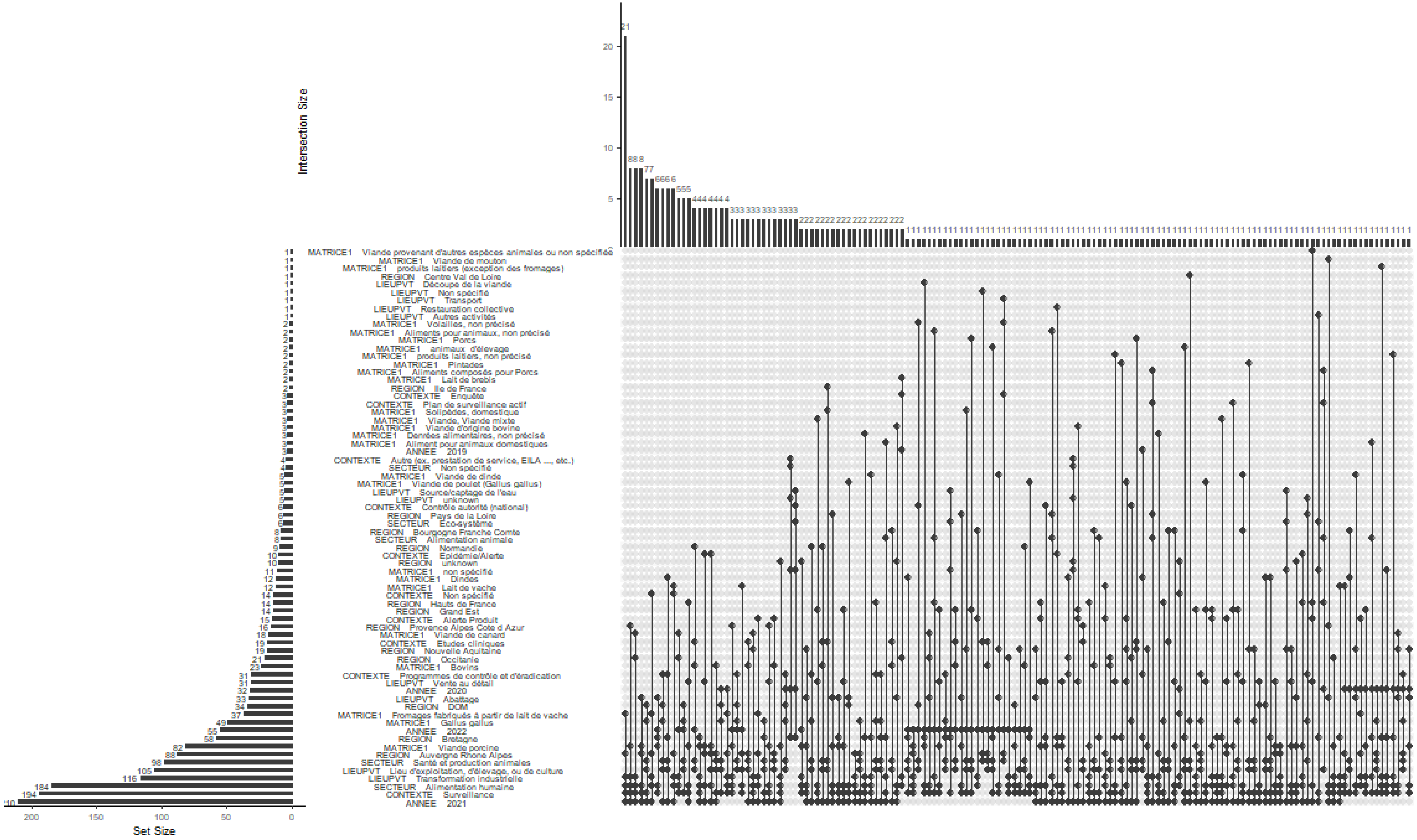
cstats_table() → optimal number of clusters



select_silhouette() → Random selection of strains in clusters

Strain select R package: Example of application

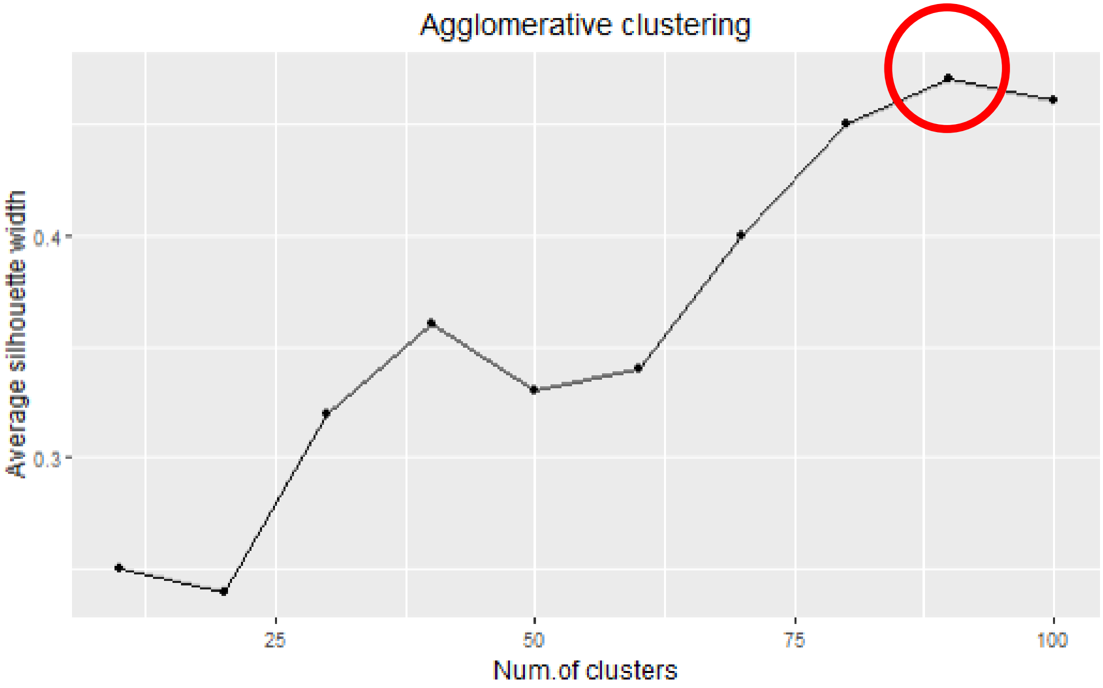
Dataset of 300 *Salmonella* Typhimurium strains – Metadata: 24 variables
→ 6 variables selected with prepare_input() + ❶ Intersection-based method



→ 149 different profiles are identified

Strain select R package: Example of application

Dataset of 300 *Salmonella* Typhimurium strains – Metadata: 24 variables
→ 6 variables selected with prepare_input() + ❷ Clustering method (Gower)



User constraints = no more than 100

→ Optimal number of clusters = 90

90 strains selected tagged in the output

- Metadata (reminder of input)
- Group (n° 1:90)
- Selection (Y/N)

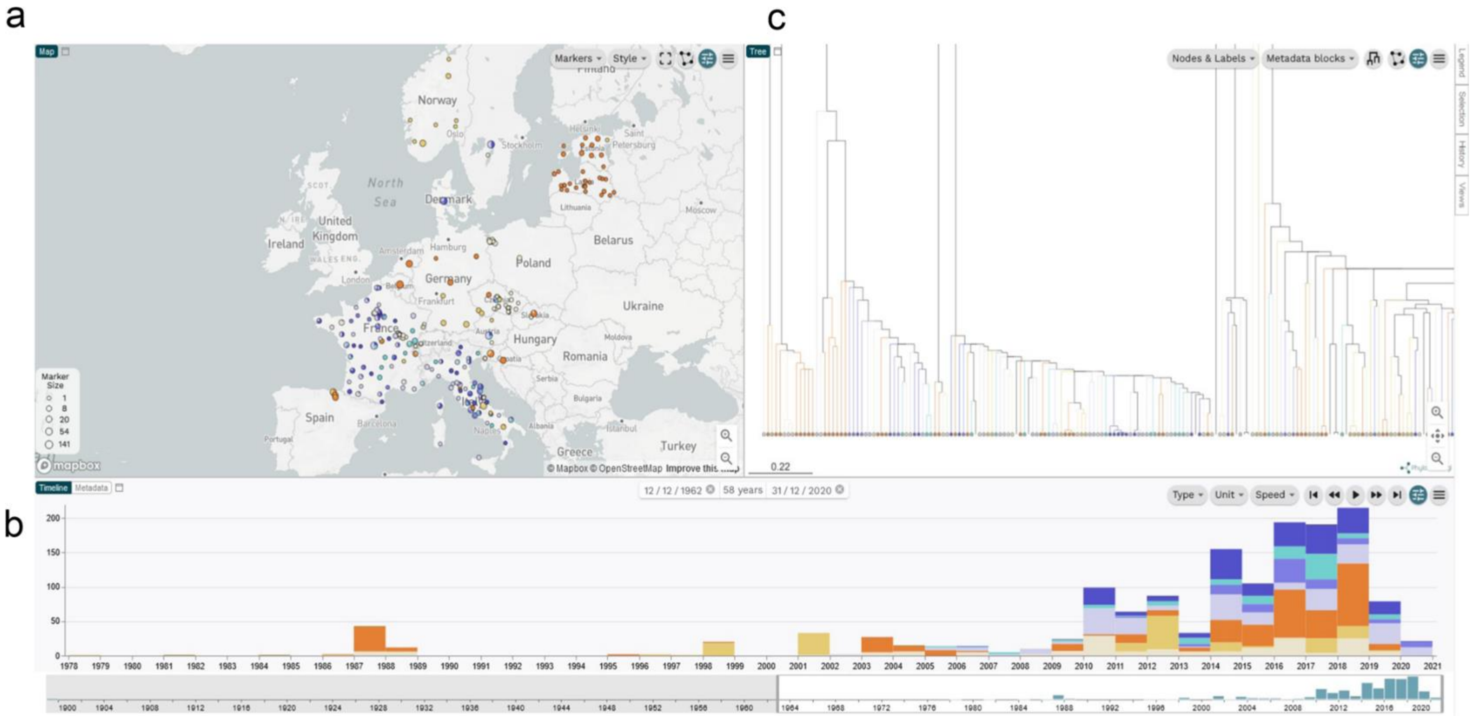
Mystrains_selection.xlsx

Projects that have used Strain select method

Selection of environmental *L. monocytogenes* strains

Felix et al. (2022)

https://doi.org/10.1038/s41597-022-01278-6

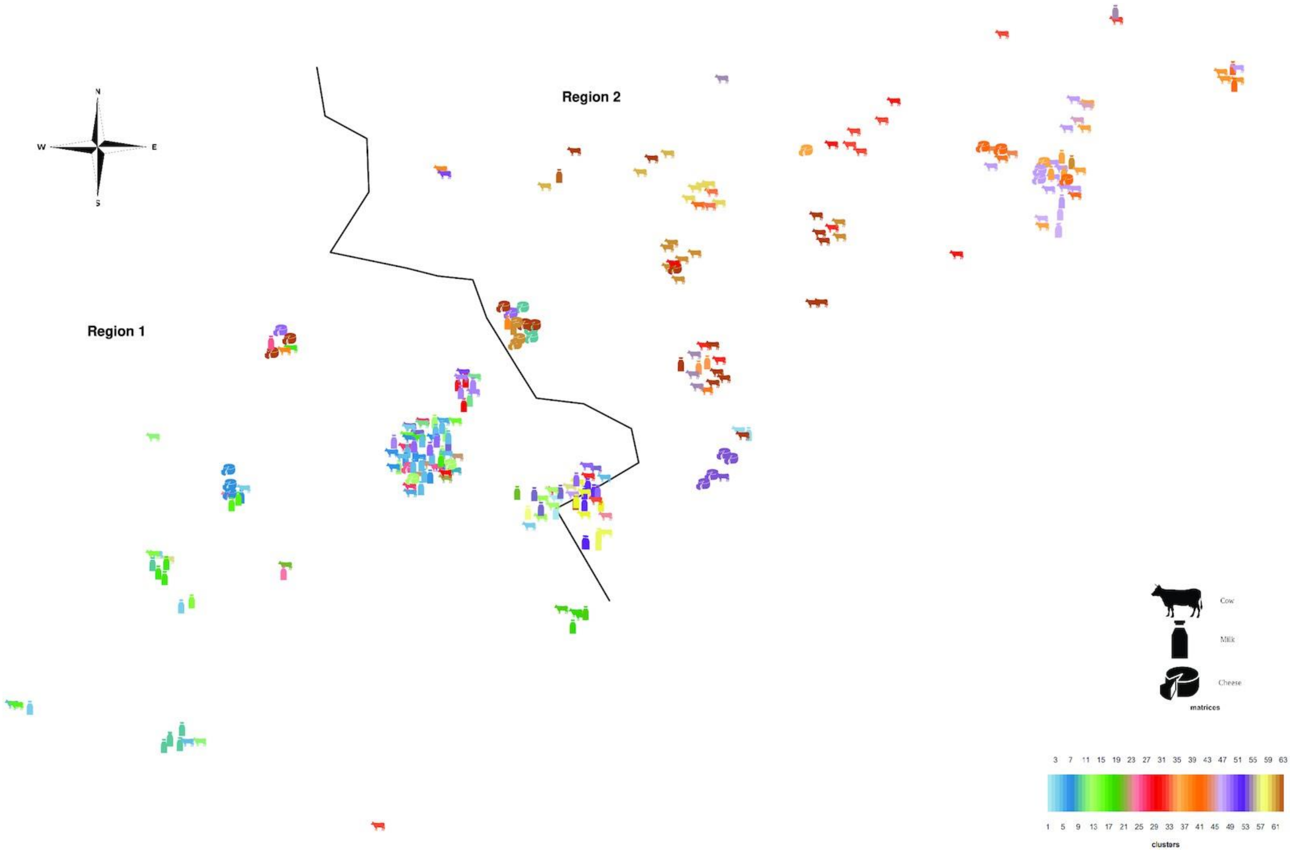


Projects that have used Strain select method

Selection on *Salmonella* Dublin strains: population structure and epidemiological investigations (France)

De Sousa Violante et al. (2022)

https://doi.org/10.1093/nargab/lqac047



Additional informations on Strain select

Availability: https://github.com/valleemarie/strainselect_package

Publication (2023) submission to https://open-research-europe.ec.europa.eu/browse/articles

Follow on activities

- Continuation of the work of the CARE project:
  - Method to select strains within CARE collection
  - Method to integrate potential new strains
- Support to EU-RLs: feel free to contact laurent.guillier@anses.fr

Marie Vallée

Thank you for your attention!

This presentation is part of the European Joint Programme One Health EJP. This programme has received funding from the European Union’s Horizon 2020 research and innovation programme under Grant Agreement No 773830.