



Bacterial strain Taxonomy for Genomic Surveillance

# ReportTree: Clustering and Visualization for Surveillance (practical)

October 23<sup>rd</sup>, 2025

# Intended Learning Objectives

Specific objectives of this session:

1. Learn what is ReporTree and its utility in routine genomic surveillance
2. Learn the principles behind ReporTree cluster nomenclature
3. Learn how clustering stability regions can be used for nomenclature design
4. Learn how to launch a ReporTree command line for routine surveillance

# Outline

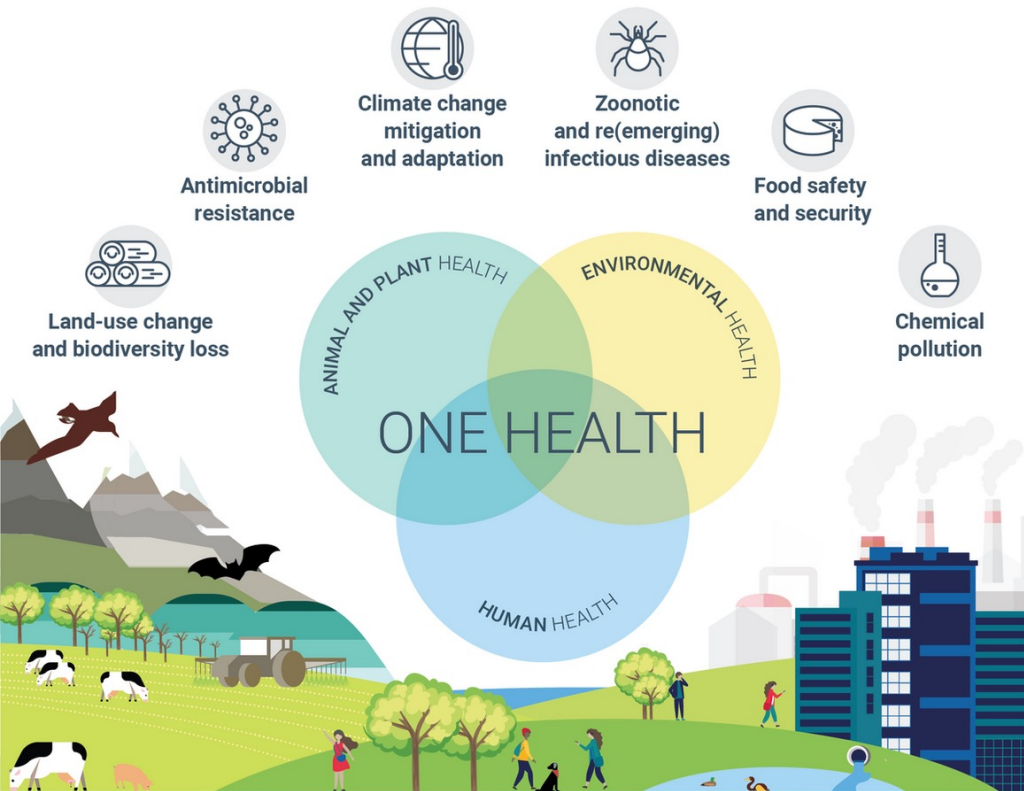
This session consists of the following elements:

1. Introduction to ReporTree
2. ReporTree's input flexibility and main outputs
3. Detect and track genetic clusters with ReporTree nomenclature system
4. Identification of candidate thresholds for nomenclature design (clustering stability regions)
5. Exercise of the real-life application of ReporTree in routine surveillance

# Genomic surveillance

*Genomic surveillance is the process of constantly monitoring pathogens and analyzing their genetic similarities and differences.*

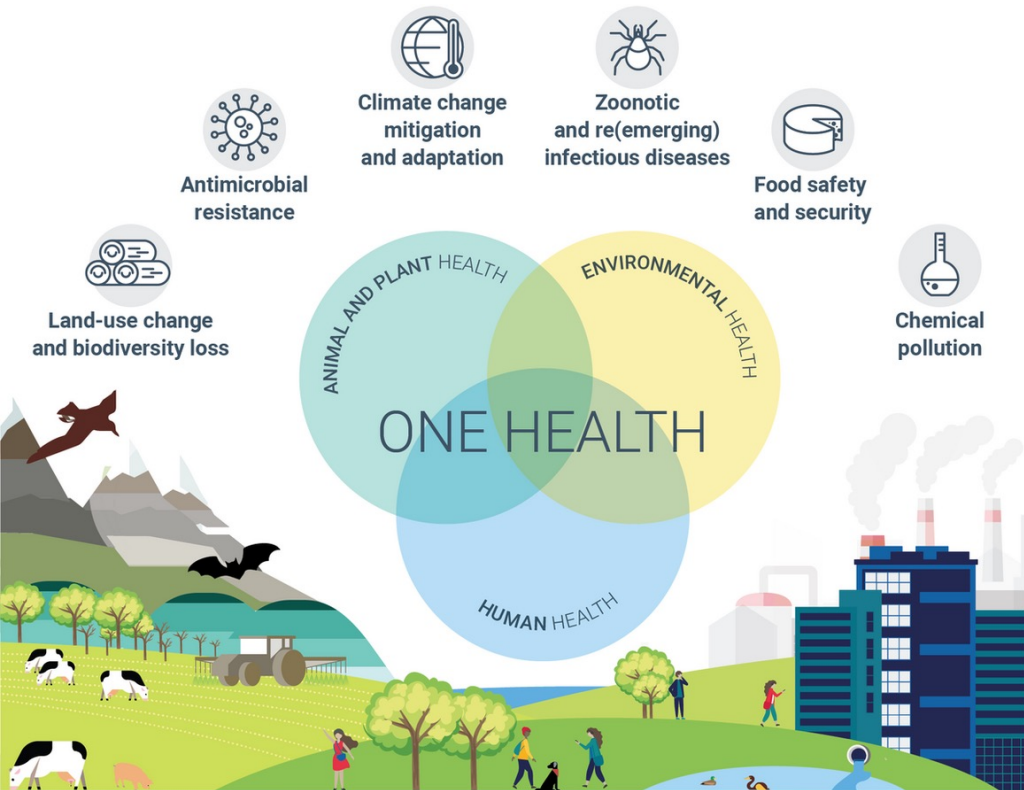
WHO, 2022



# Genomic surveillance

*Genomic surveillance is the process of constantly monitoring pathogens and analyzing their genetic similarities and differences.*

WHO, 2022



A pivotal outcome of genomics surveillance is the **identification of pathogen genetic clusters/lineages** and their **characterization** in terms of geotemporal spread or linkage to clinical and demographic data

# Analyses of Whole-Genome Sequencing data

Multiple bioinformatics solutions, but similar steps and goals...

## Genetic data

|   |   |   |   |   |   |   |   |
|---|---|---|---|---|---|---|---|
| A | T | C | G | A | T | C | G |
| A | T | C | G | A | T | C | C |
| N | A | C | G | A | T | C | C |
| A | C | G | G | A | T | C | C |
| N | C | G | N | A | T | N | C |

SNPs\* or alleles

\*Single-Nucleotide Polymorphisms

# Analyses of Whole-Genome Sequencing data

Multiple bioinformatics solutions, but similar steps and goals...

## Genetic data

|   |   |   |   |   |   |   |   |
|---|---|---|---|---|---|---|---|
| A | T | C | G | A | T | C | G |
| A | T | C | G | A | T | C | C |
| N | A | C | G | A | T | C | C |
| A | C | G | G | A | T | C | C |
| N | C | G | N | A | T | N | C |

SNPs\* or alleles

Pairwise distance  
between samples

# Analyses of Whole-Genome Sequencing data

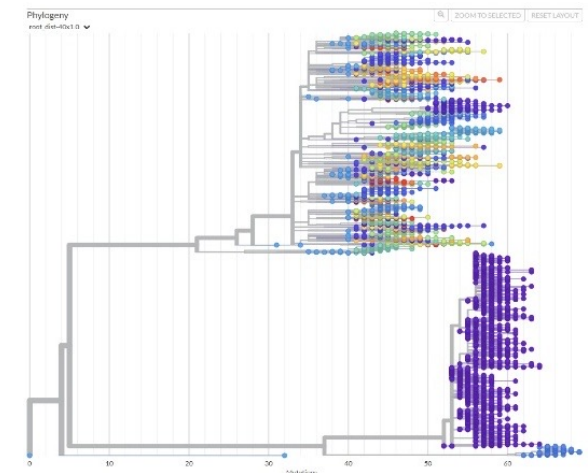
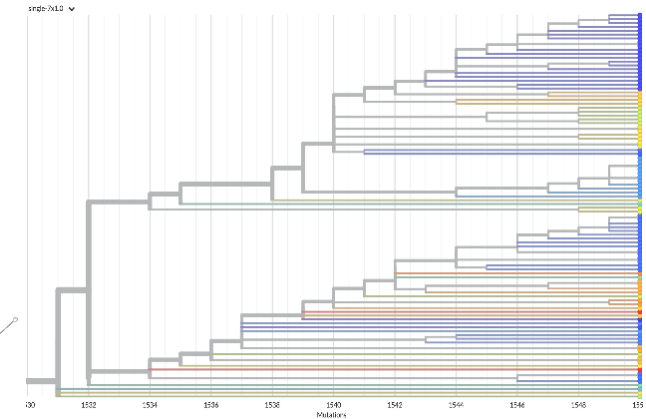
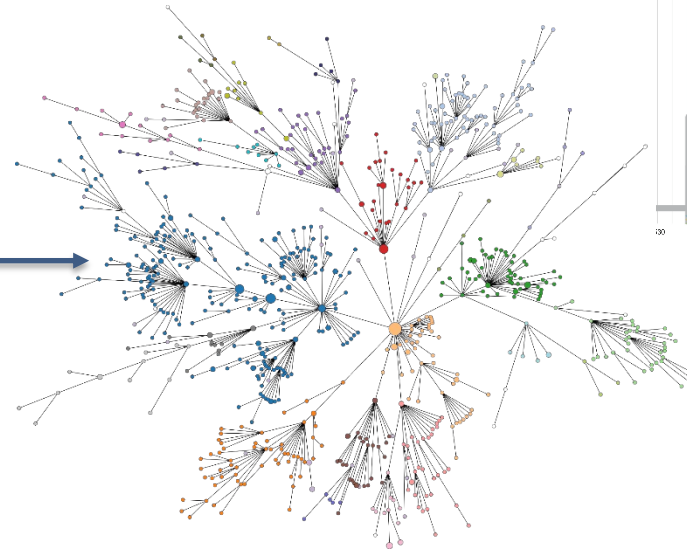
Multiple bioinformatics solutions, but similar steps and goals...

## Genetic data

|   |   |   |   |   |   |   |   |
|---|---|---|---|---|---|---|---|
| A | T | C | G | A | T | C | G |
| A | T | C | G | A | T | C | C |
| N | A | C | G | A | T | C | C |
| A | C | G | G | A | T | C | C |
| N | C | G | N | A | T | N | C |

SNPs\* or alleles

Pairwise distance  
between samples



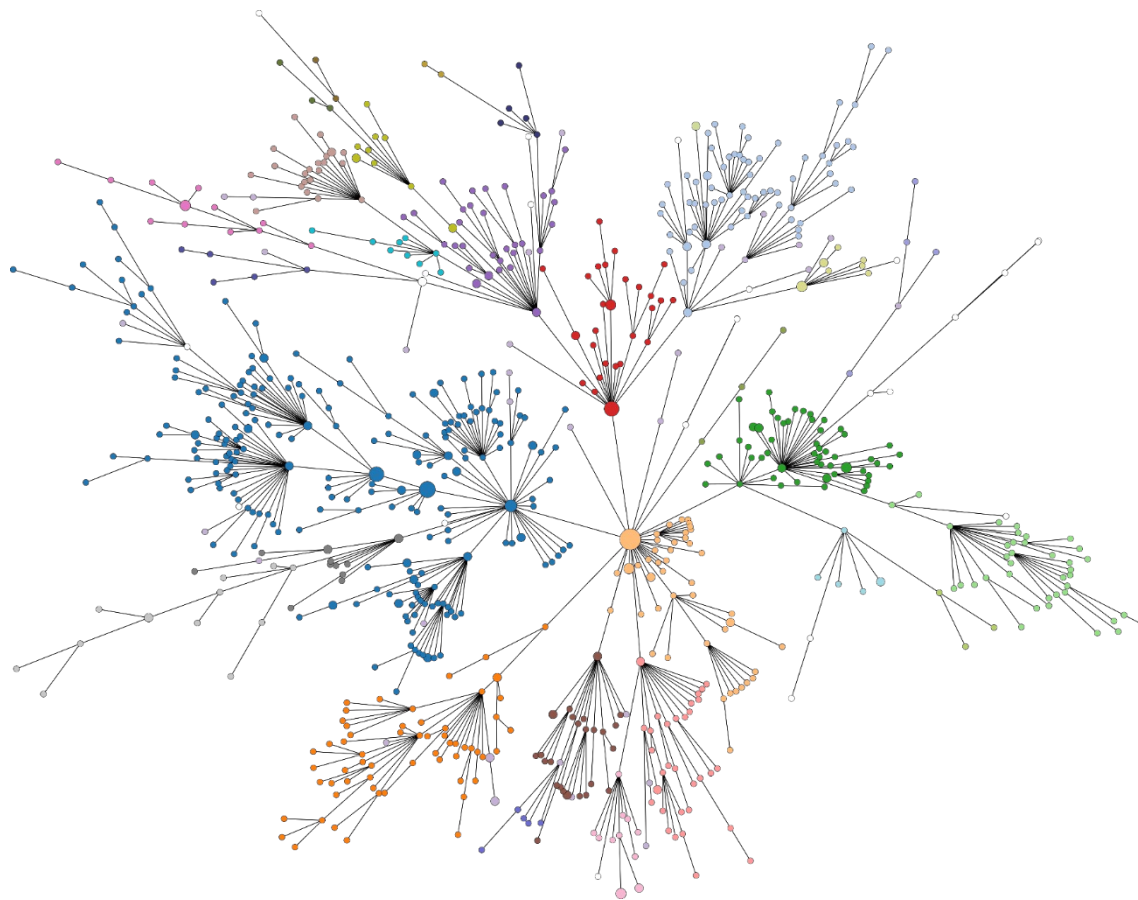
\*Single-Nucleotide Polymorphisms





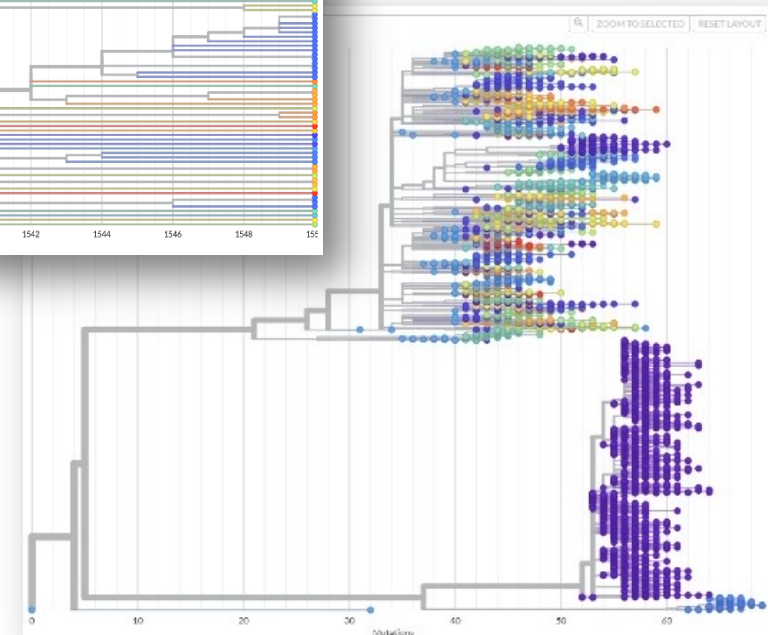
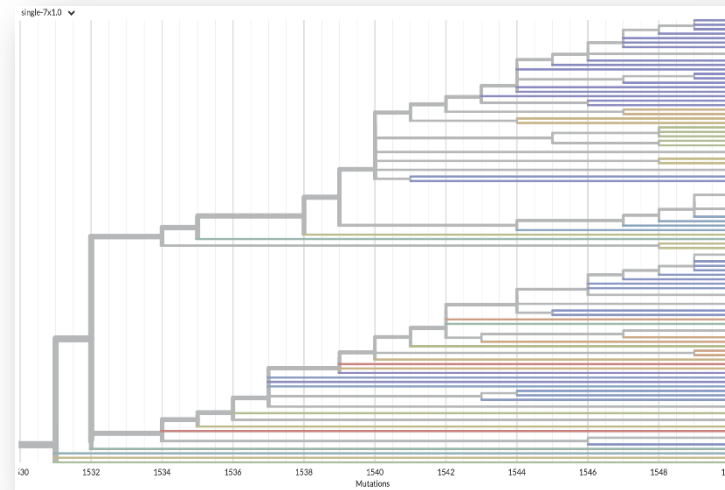
# How do you perform bacterial clustering analyses?

# Detection of genetic clusters and characterization

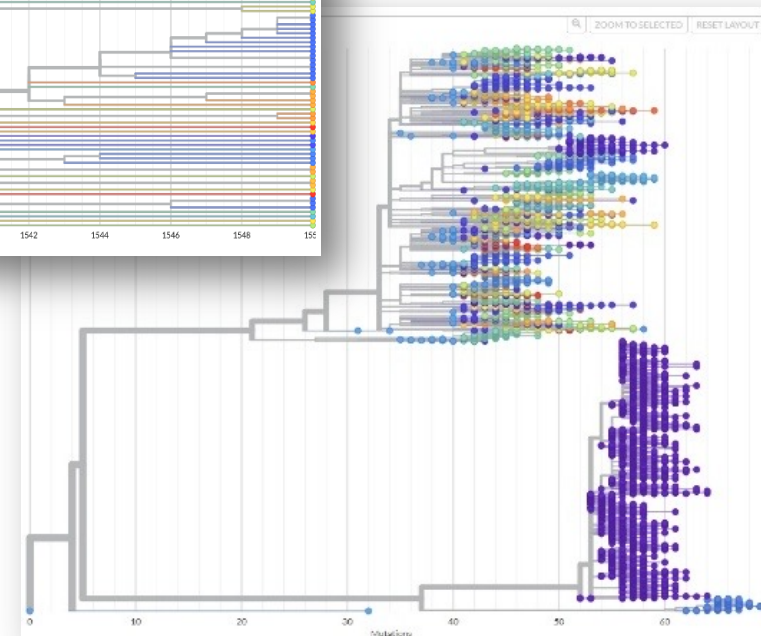
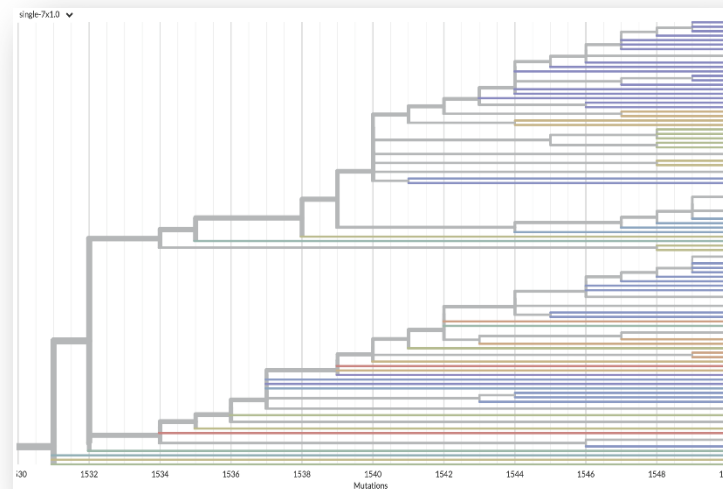
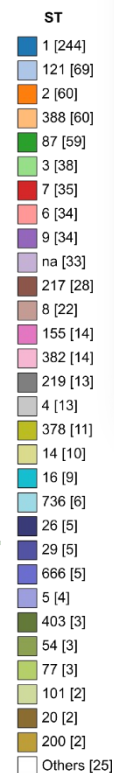
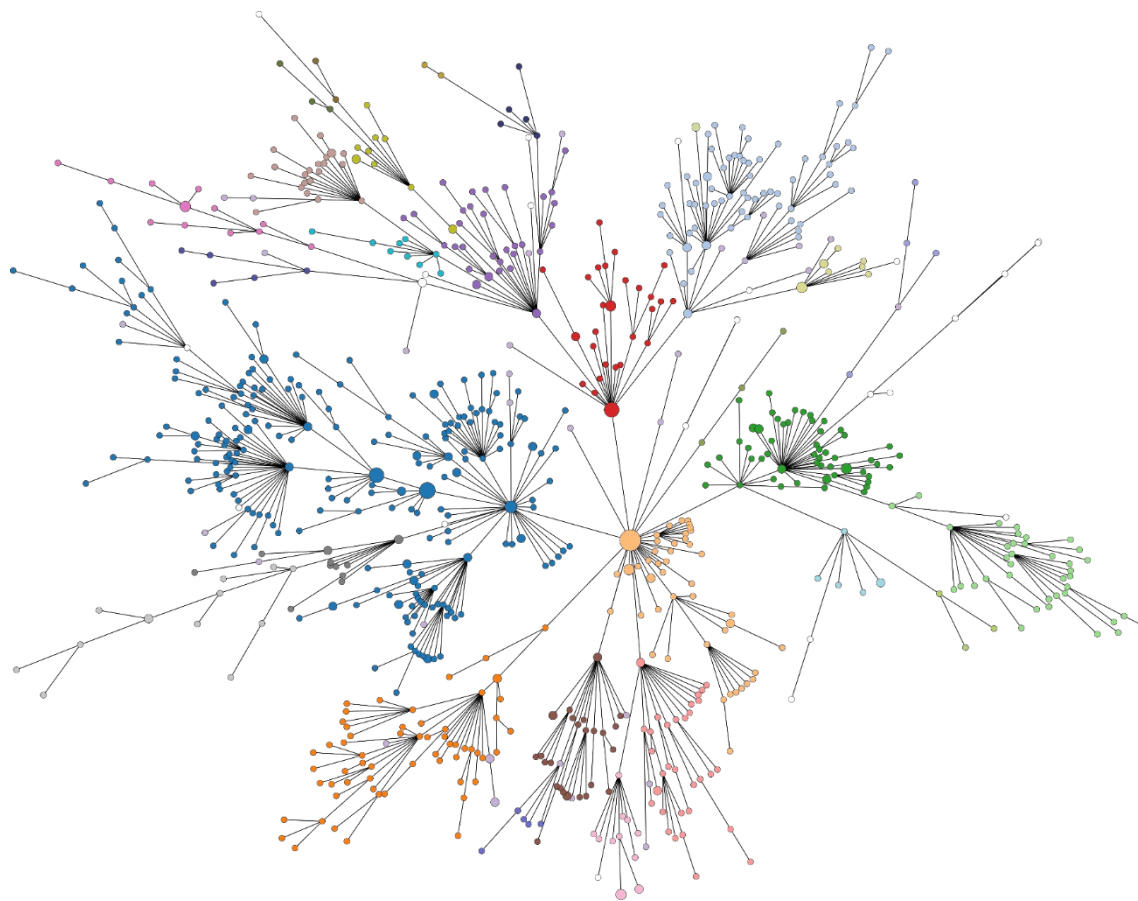


ST

|        |       |
|--------|-------|
| 1      | [244] |
| 121    | [69]  |
| 2      | [60]  |
| 388    | [60]  |
| 87     | [59]  |
| 3      | [38]  |
| 7      | [35]  |
| 6      | [34]  |
| 9      | [34]  |
| na     | [33]  |
| 217    | [28]  |
| 8      | [22]  |
| 155    | [14]  |
| 382    | [14]  |
| 219    | [13]  |
| 4      | [13]  |
| 378    | [11]  |
| 14     | [10]  |
| 16     | [9]   |
| 736    | [6]   |
| 26     | [5]   |
| 29     | [5]   |
| 666    | [5]   |
| 5      | [4]   |
| 403    | [3]   |
| 54     | [3]   |
| 77     | [3]   |
| 101    | [2]   |
| 20     | [2]   |
| 200    | [2]   |
| Others | [25]  |



# Detection of genetic clusters and characterization



*The **detection of genetic clusters and their characterization** with clinical/epidemiological data is still a challenging step that often relies on **non-automated** approaches*

# ReporTree

- Flexible solution to automatically identify genetic clusters at any (or all) distance thresholds
- Generate surveillance-oriented reports based on the available metadata, such as timespan, geography or clinical status



# ReporTree

- Flexible solution to automatically identify genetic clusters at any (or all) distance thresholds
- Generate surveillance-oriented reports based on the available metadata, such as timespan, geography or clinical status



Software | [Open access](#) | Published: 15 June 2023

**ReporTree: a surveillance-oriented tool to strengthen the linkage between pathogen genetic clusters and epidemiological data**

[Verónica Mixão](#), [Miguel Pinto](#), [Daniel Sobral](#), [Adriano Di Pasquale](#), [João Paulo Gomes](#) & [Vítor Borges](#) 

[Genome Medicine](#) **15**, Article number: 43 (2023) | [Cite this article](#)

**4644** Accesses | **24** Citations | **18** Altmetric | [Metrics](#)



# ReporTree

- Flexible solution to automatically identify genetic clusters at any (or all) distance thresholds
- Generate surveillance-oriented reports based on the available metadata, such as timespan, geography or clinical status



Software | [Open access](#) | Published: 15 June 2023

**ReporTree: a surveillance-oriented tool to strengthen the linkage between pathogen genetic clusters and epidemiological data**

[Verónica Mixão](#), [Miguel Pinto](#), [Daniel Sobral](#), [Adriano Di Pasquale](#), [João Paulo Gomes](#) & [Vítor Borges](#) 

[Genome Medicine](#) **15**, Article number: 43 (2023) | [Cite this article](#)

**4644** Accesses | **24** Citations | **18** Altmetric | [Metrics](#)



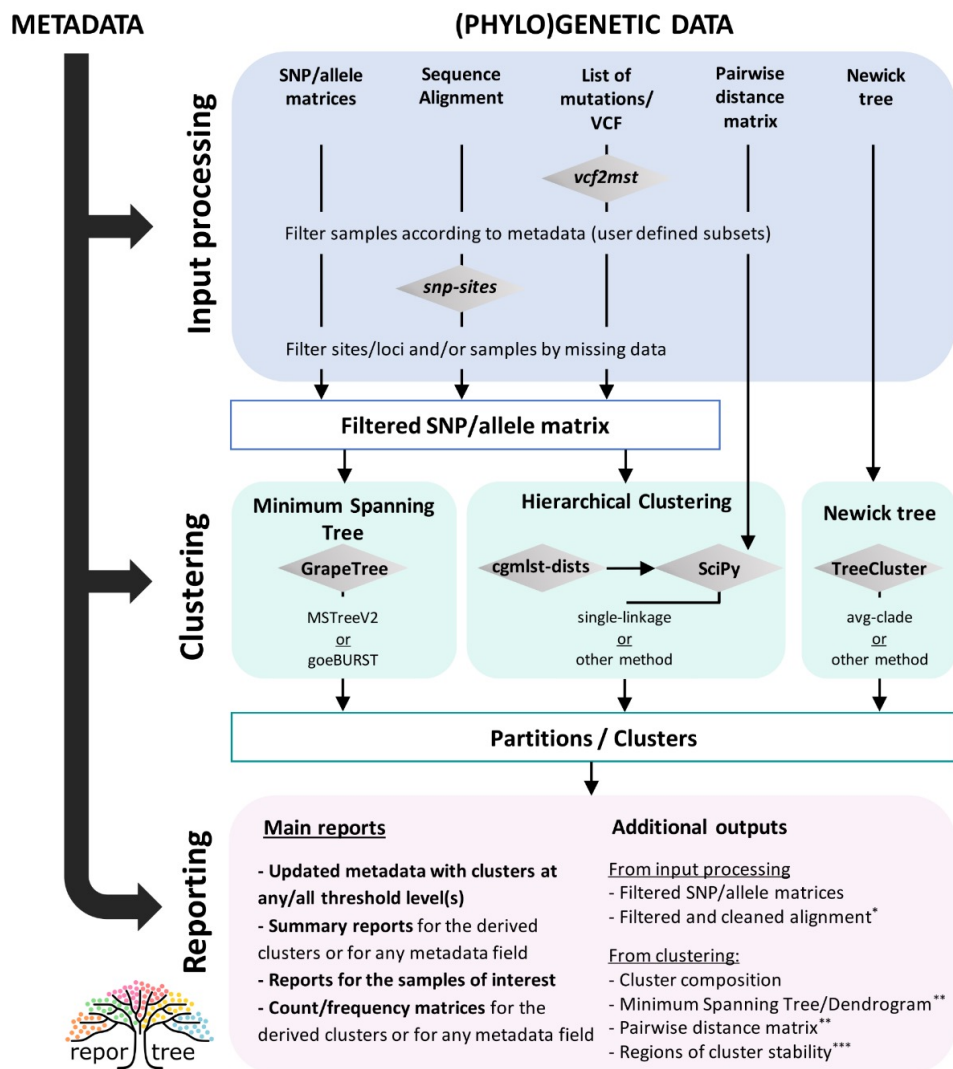
Also available through:  
**PubMLST, COHESIVE, etc.**



**Which WGS-based typing approach do you use for routine surveillance of bacterial pathogens?**

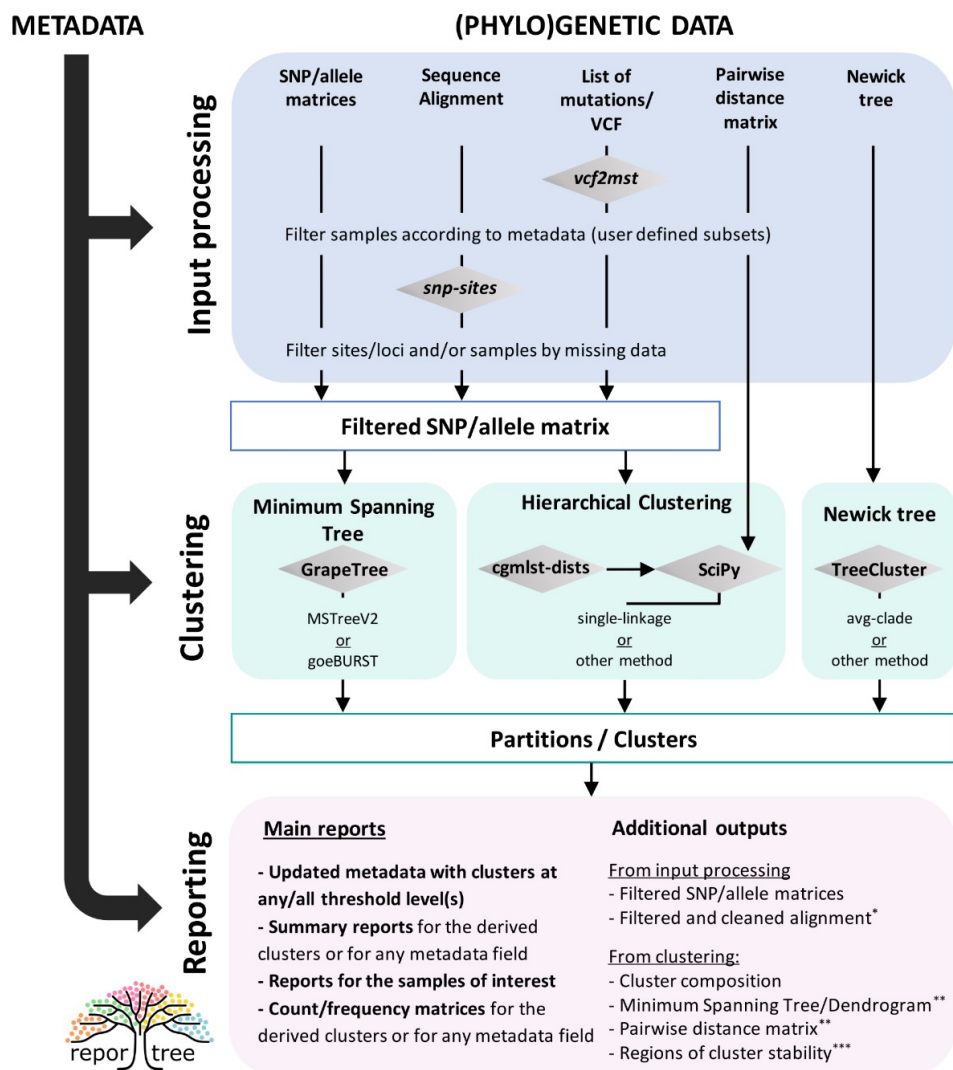


# Designed for surveillance of multiple pathogens





# Designed for surveillance of multiple pathogens



## Main outputs:

- Tree
- Metadata table updated with clusters at any/all resolution levels
- Summary reports with the statistics/trends for the derived genetic clusters
- Summary reports for the samples of interest (e.g., new samples)

| partition  | cluster     | cluster_length | samples                             | source                         | country                         | first_seq_date | last_seq_date |
|------------|-------------|----------------|-------------------------------------|--------------------------------|---------------------------------|----------------|---------------|
| MST-4x1.0  | cluster_34  | 2              | sample_0419,sample_0464             | clinical (50.0%), food (50.0%) | C (50.0%), B (50.0%) (n = 2)    | 20/06/11       | 12/03/12      |
| MST-7x1.0  | cluster_106 | 33             | sample_0017,sample_0037,sample_0037 | clinical (63.6%), food (36.4%) | C (42.4%), B (30.3%), A (27.3%) | 30/03/04       | 14/07/21      |
| MST-14x1.0 | cluster_87  | 46             | sample_0017,sample_0037,sample_0037 | clinical (67.4%), food (32.6%) | A (37.0%), C (34.8%), B (28.3%) | 30/03/04       | 14/07/21      |

# How does it work?

# ReporTree workflow

## Genetic data

|   |   |   |   |   |   |   |   |
|---|---|---|---|---|---|---|---|
| A | T | C | G | A | T | C | G |
| A | T | C | G | A | T | C | C |
| N | A | C | G | A | T | C | C |
| A | C | G | G | A | T | C | C |
| N | C | G | N | A | T | N | C |

SNPs or alleles

# ReporTree workflow

## Genetic data

|   |   |   |   |   |   |   |   |
|---|---|---|---|---|---|---|---|
| A | T | C | G | A | T | C | G |
| A | T | C | G | A | T | C | C |
| N | A | C | G | A | T | C | C |
| A | C | G | G | A | T | C | C |
| N | C | G | N | A | T | N | C |

SNPs or alleles



Treated similar to wgMLST data  
(*alignment\_processing.py*)

# ReporTree workflow

## Genetic data

|   |   |   |   |   |   |   |   |
|---|---|---|---|---|---|---|---|
| A | T | C | G | A | T | C | G |
| A | T | C | G | A | T | C | C |
| N | A | C | G | A | T | C | C |
| A | C | G | G | A | T | C | C |
| N | C | G | N | A | T | N | C |

SNPs or alleles

cgMLST  
(set of fixed core loci)

wgMLST  
(set of core loci +  
accessory loci)



# ReportTree workflow

## Genetic data

|   |   |   |   |   |   |   |   |
|---|---|---|---|---|---|---|---|
| A | T | C | G | A | T | C | G |
| A | T | C | G | A | T | C | C |
| N | A | C | G | A | T | C | C |
| A | C | G | G | A | T | C | C |
| N | C | G | N | A | T | N | C |

SNPs or alleles

cgMLST  
(set of fixed core loci)

wgMLST  
(set of core loci +  
accessory loci)

remove loci/positions  
with excess of missing data

remove samples  
with excess of missing data



# ReporTree workflow

## Genetic data

|   |   |   |   |   |   |   |   |
|---|---|---|---|---|---|---|---|
| A | T | C | G | A | T | C | G |
| A | T | C | G | A | T | C | C |
| N | A | C | G | A | T | C | C |
| A | C | G | G | A | T | C | C |
| N | C | G | N | A | T | N | C |

SNPs or alleles

cgMLST  
(set of fixed core loci)

wgMLST  
(set of core loci +  
accessory loci)

remove loci/positions  
with excess of missing data

remove samples  
with excess of missing data

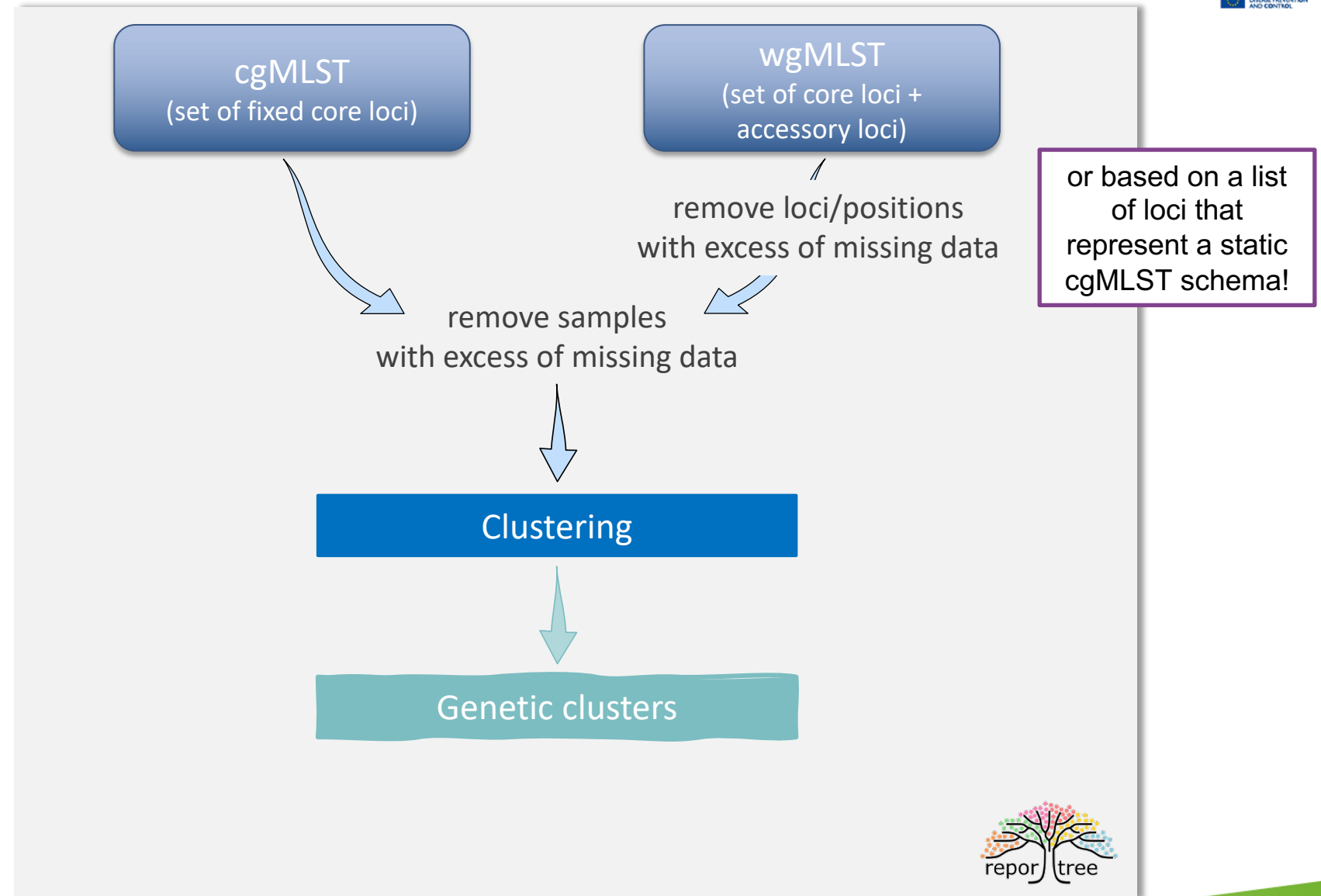
or based on a list  
of loci that  
represent a static  
cgMLST schema!

# ReporTree workflow

## Genetic data

|   |   |   |   |   |   |   |   |
|---|---|---|---|---|---|---|---|
| A | T | C | G | A | T | C | G |
| A | T | C | G | A | T | C | C |
| N | A | C | G | A | T | C | C |
| A | C | G | G | A | T | C | C |
| N | C | G | N | A | T | N | C |

SNPs or alleles



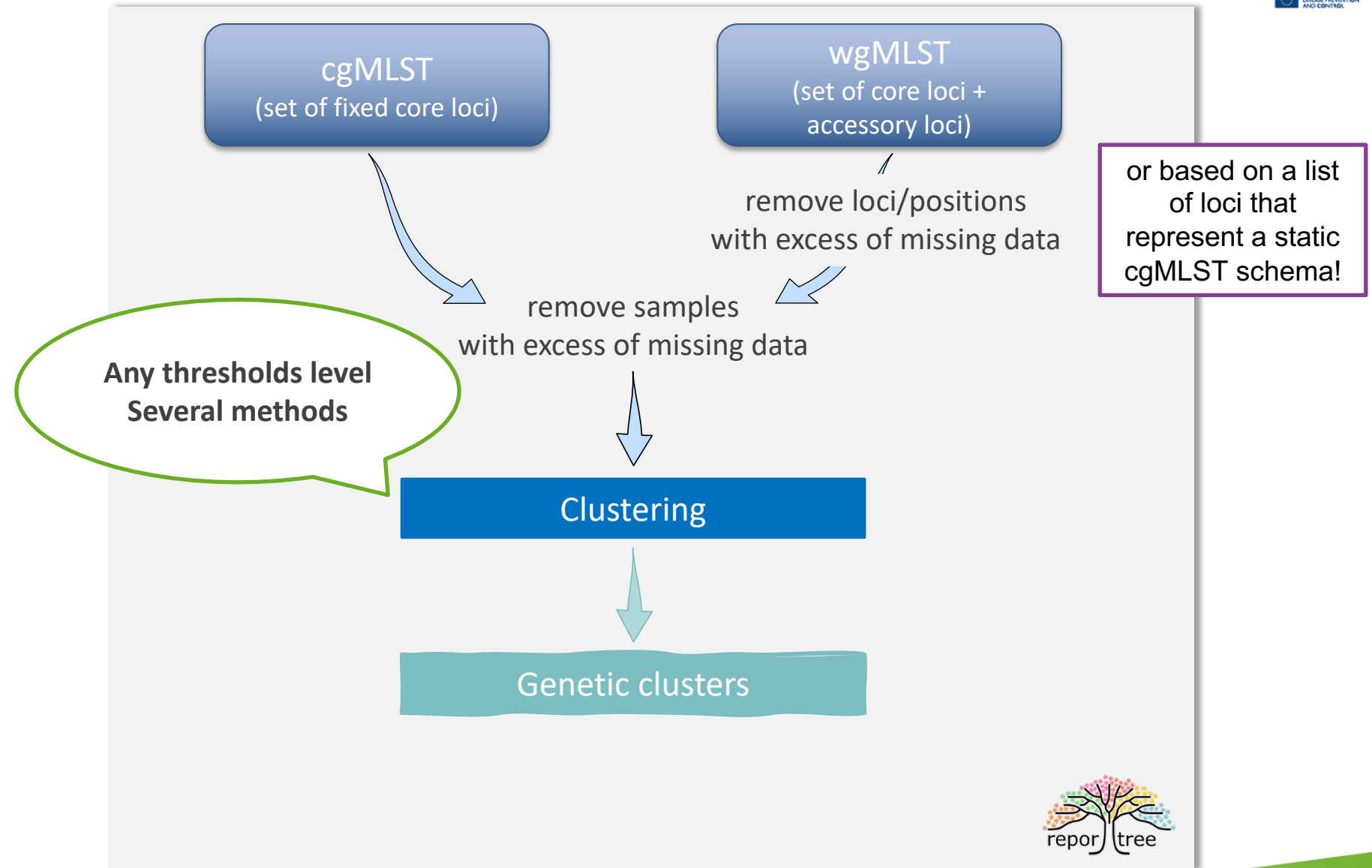


# ReporTree workflow

## Genetic data

|   |   |   |   |   |   |   |   |
|---|---|---|---|---|---|---|---|
| A | T | C | G | A | T | C | G |
| A | T | C | G | A | T | C | C |
| N | A | C | G | A | T | C | C |
| A | C | G | G | A | T | C | C |
| N | C | G | N | A | T | N | C |

SNPs or alleles

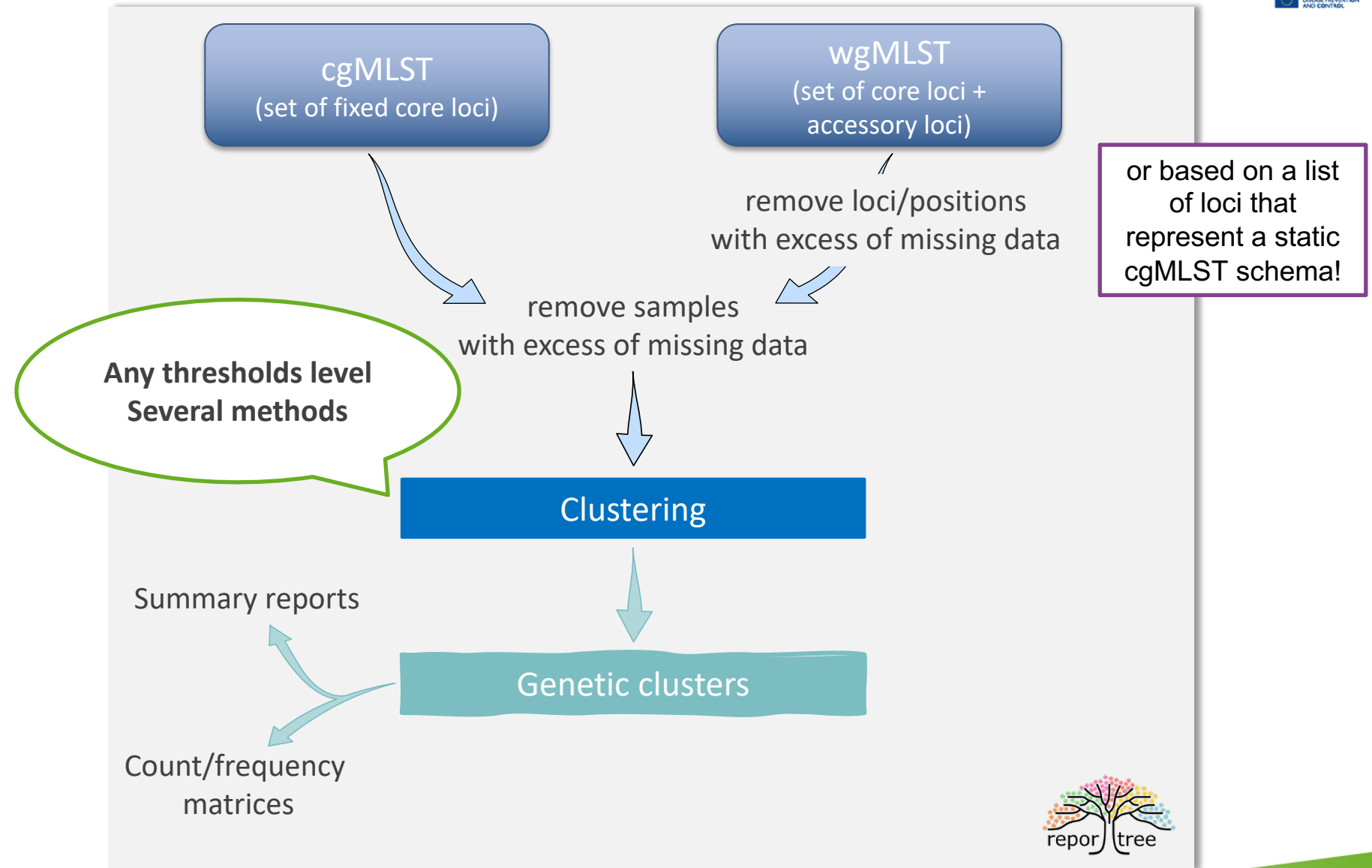


# ReporTree workflow

## Genetic data

|   |   |   |   |   |   |   |   |
|---|---|---|---|---|---|---|---|
| A | T | C | G | A | T | C | G |
| A | T | C | G | A | T | C | C |
| N | A | C | G | A | T | C | C |
| A | C | G | G | A | T | C | C |
| N | C | G | N | A | T | N | C |

SNPs or alleles

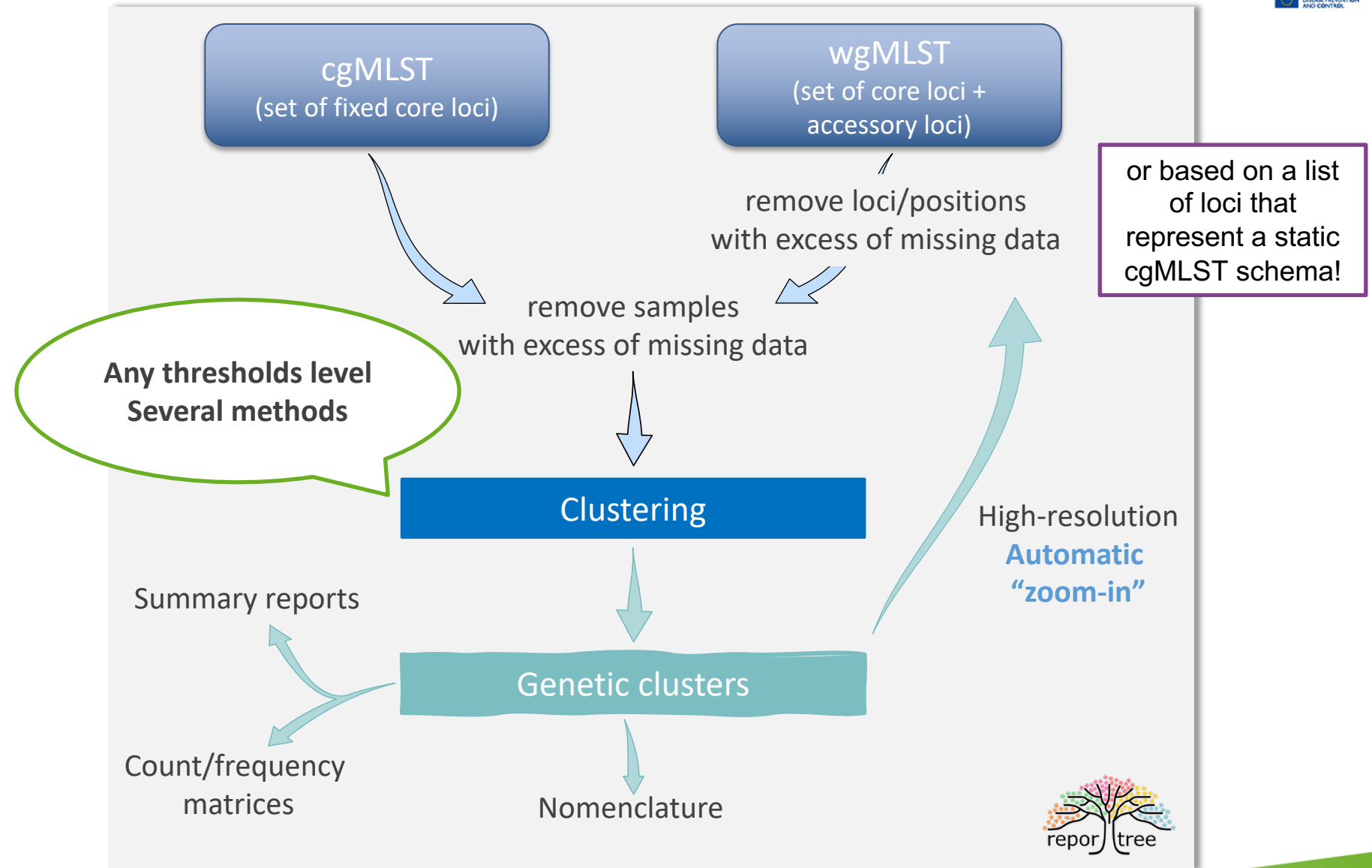


# ReportTree workflow

## Genetic data

|   |   |   |   |   |   |   |   |
|---|---|---|---|---|---|---|---|
| A | T | C | G | A | T | C | G |
| A | T | C | G | A | T | C | C |
| N | A | C | G | A | T | C | C |
| A | C | G | G | A | T | C | C |
| N | C | G | N | A | T | N | C |

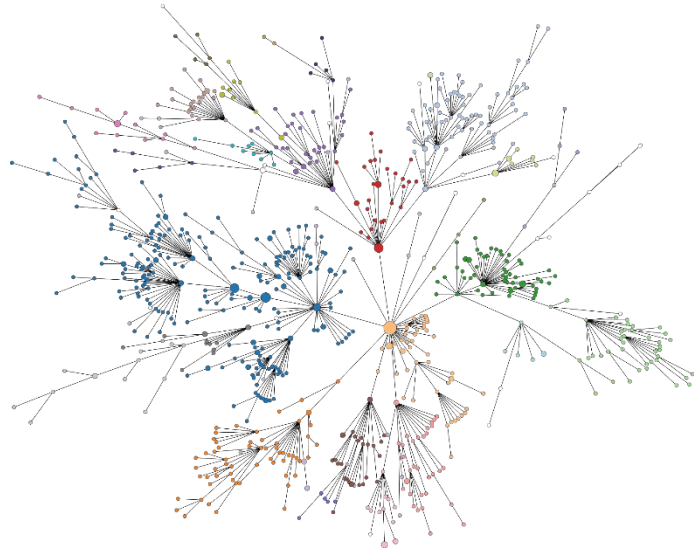
SNPs or alleles



# Dynamic wgMLST approach

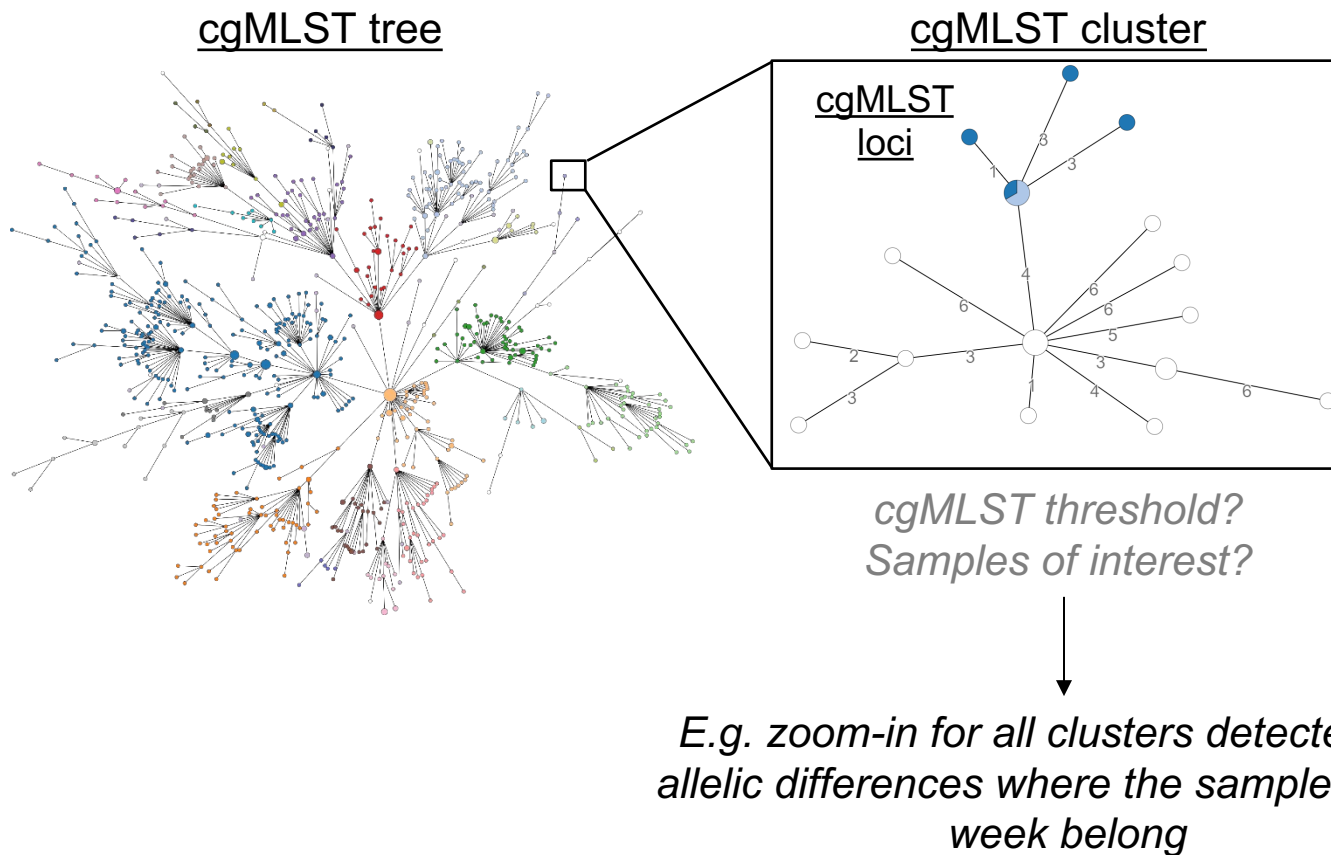
- Useful for species with wgMLST schemas (e.g. *Salmonella enterica* or *Escherichia coli*)

cgMLST tree



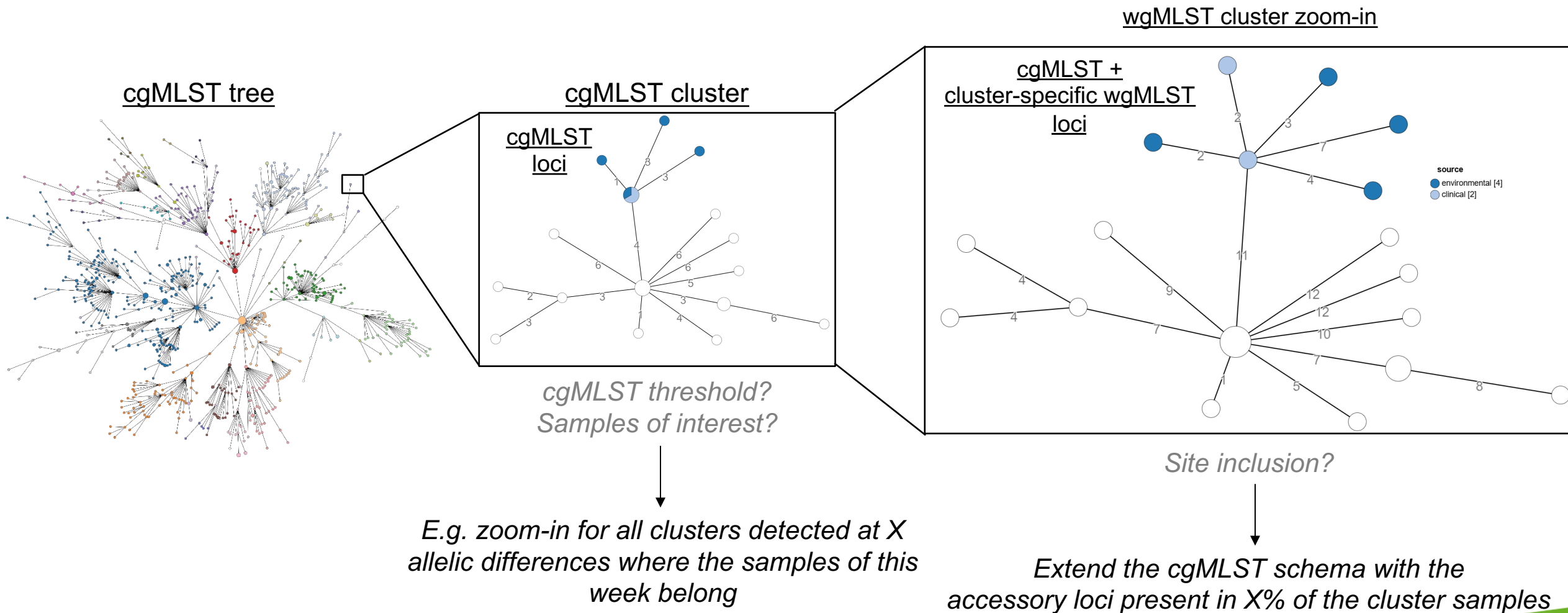
# Dynamic wgMLST approach

- Useful for species with wgMLST schemas (e.g. *Salmonella enterica* or *Escherichia coli*)



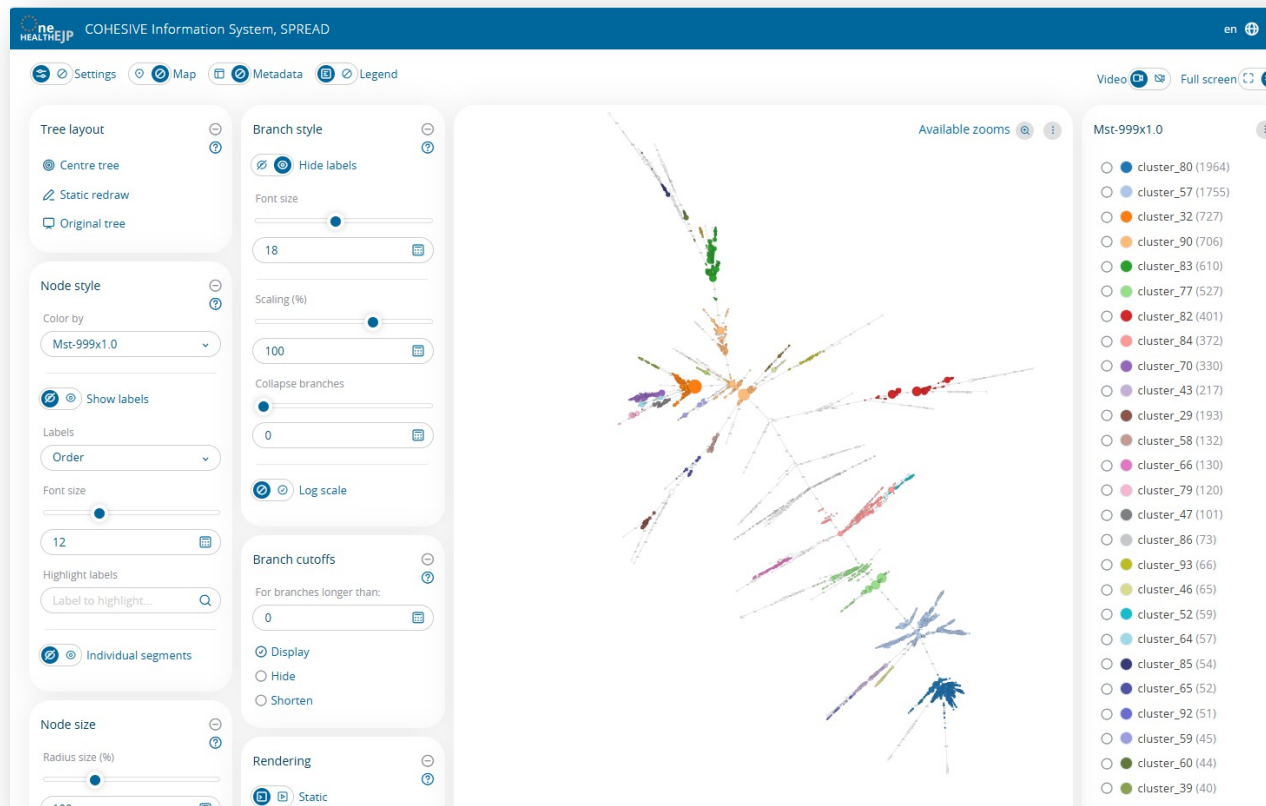
# Dynamic wgMLST approach

- Useful for species with wgMLST schemas (e.g. *Salmonella enterica* or *Escherichia coli*)



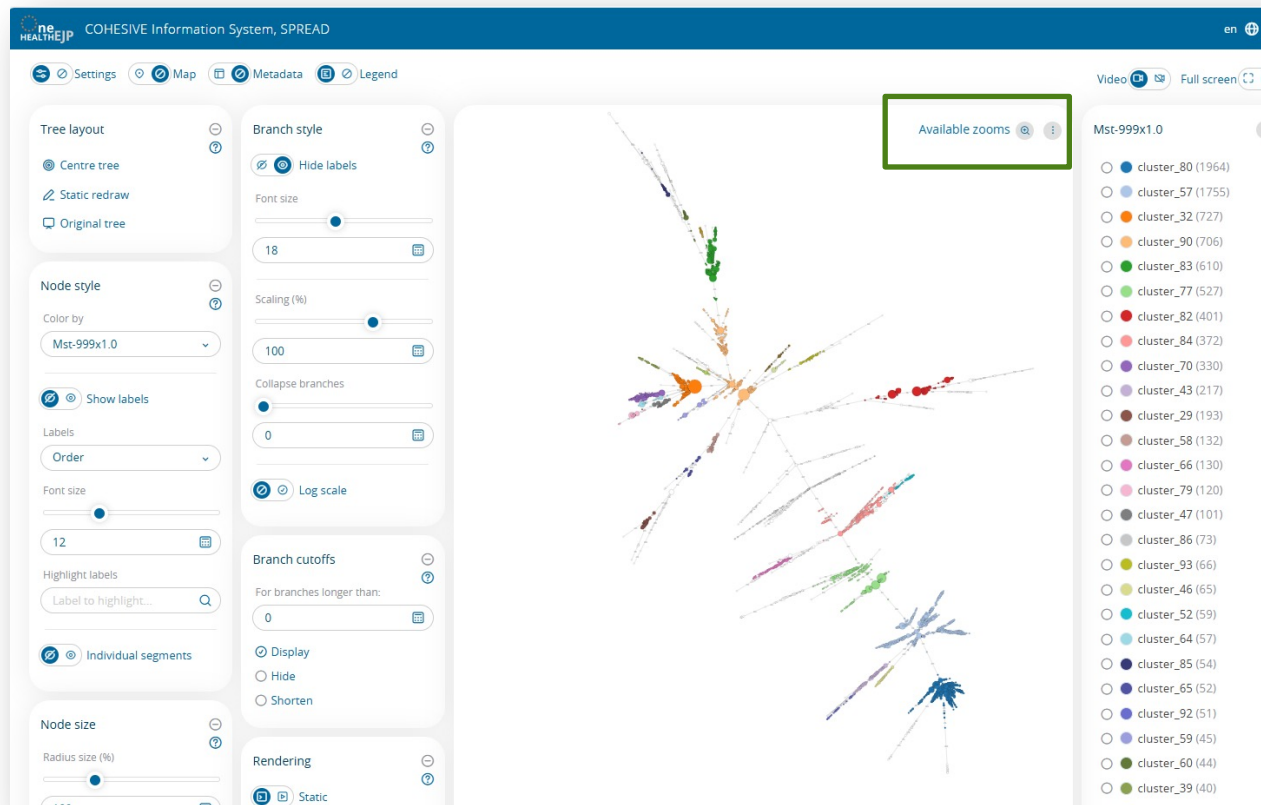
# Dynamic wgMLST approach - visualization

- SPREAD - Spatiotemporal Pathogen Relationships and Epidemiological Analysis Dashboard



# Dynamic wgMLST approach - visualization

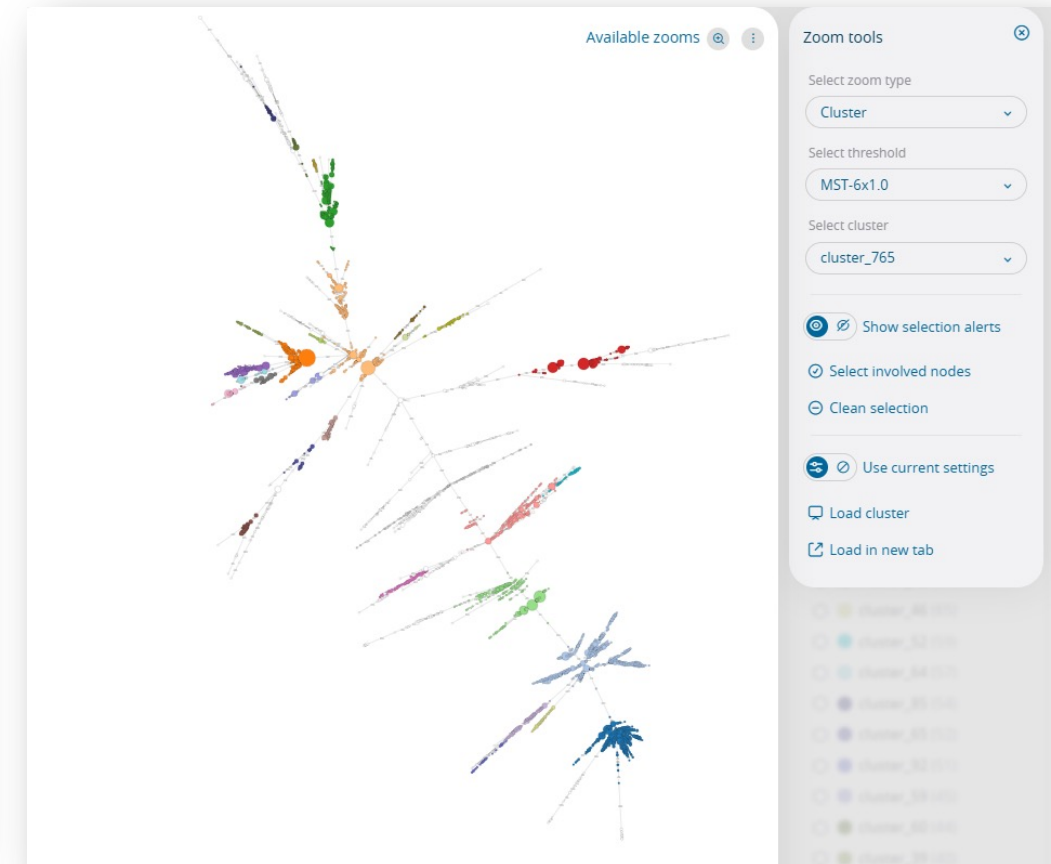
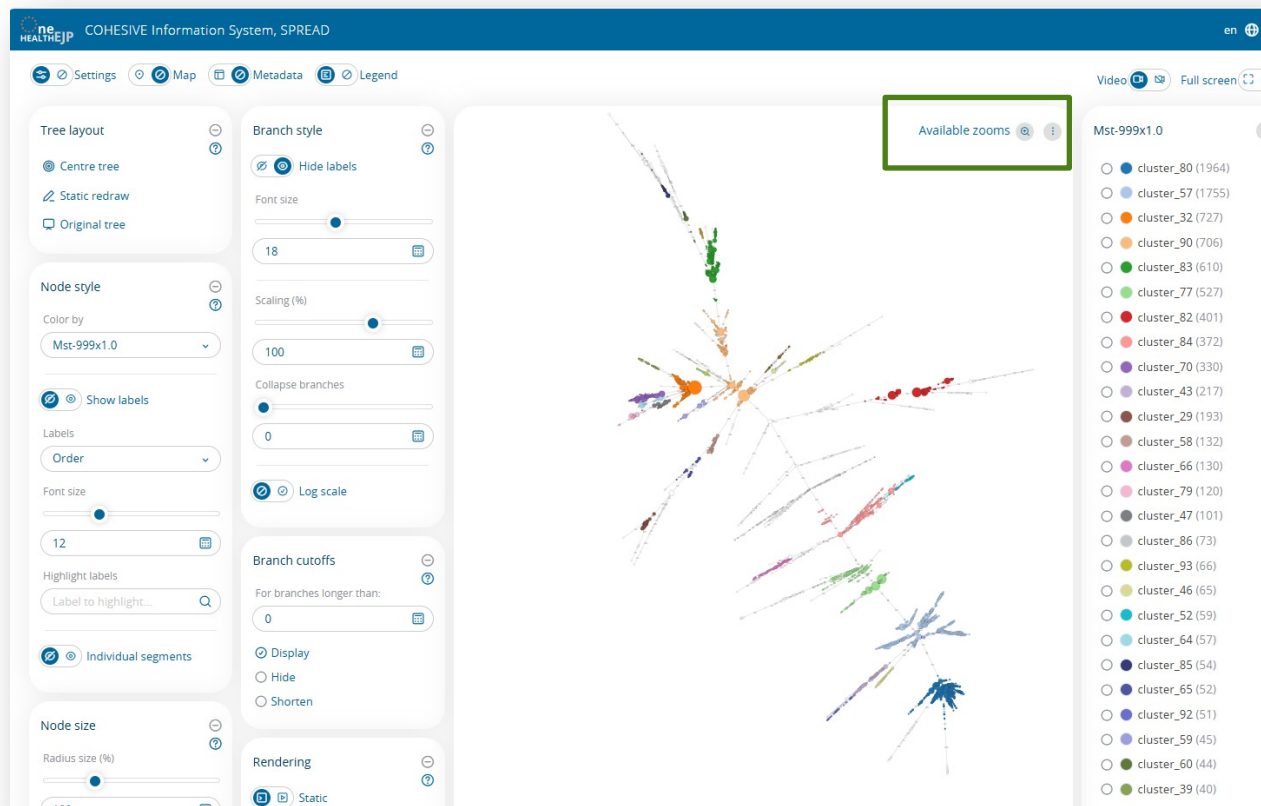
- SPREAD - Spatiotemporal Pathogen Relationships and Epidemiological Analysis Dashboard





# Dynamic wgMLST approach - visualization

- SPREAD - Spatiotemporal Pathogen Relationships and Epidemiological Analysis Dashboard

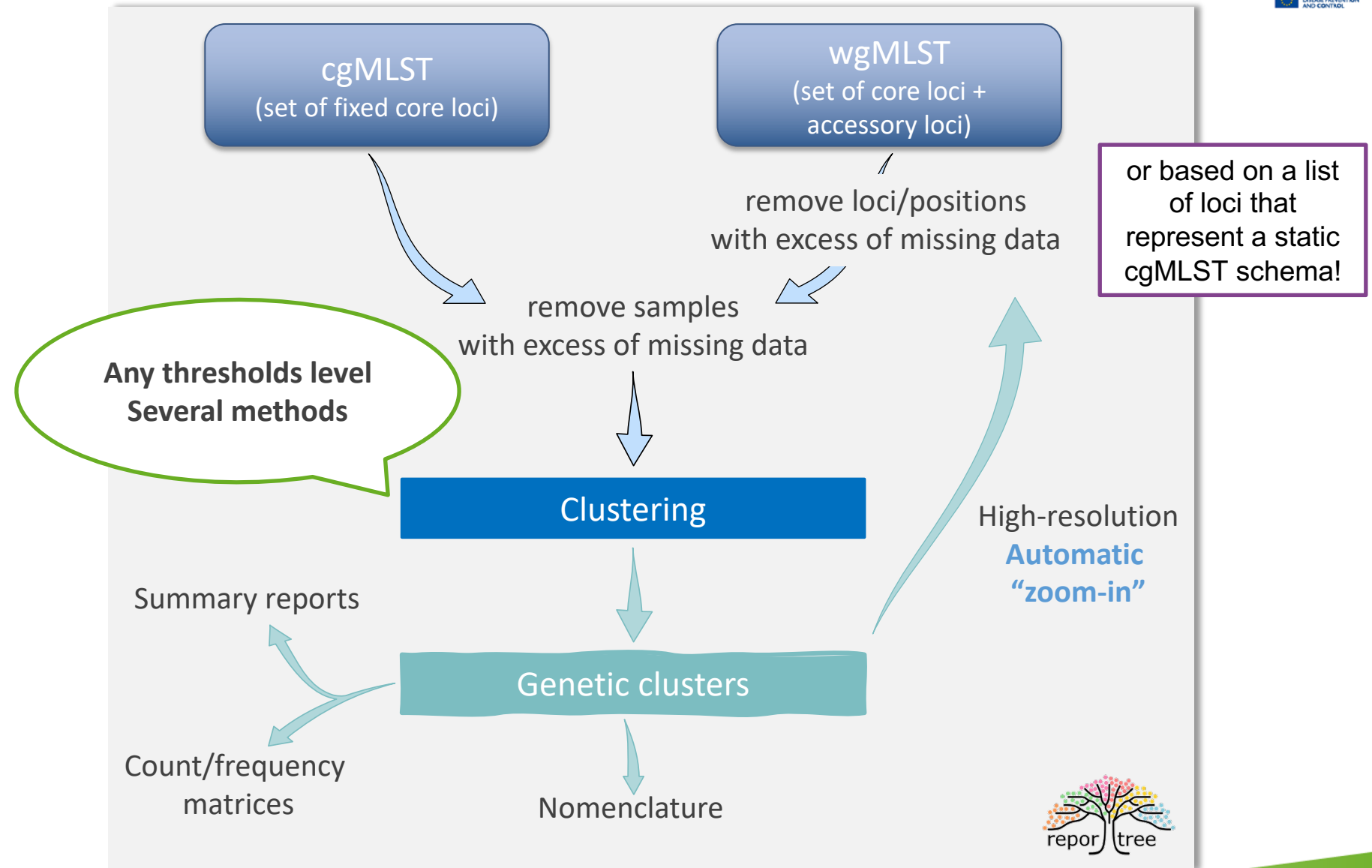


# ReporTree workflow

## Genetic data

|   |   |   |   |   |   |   |   |
|---|---|---|---|---|---|---|---|
| A | T | C | G | A | T | C | G |
| A | T | C | G | A | T | C | C |
| N | A | C | G | A | T | C | C |
| A | C | G | G | A | T | C | C |
| N | C | G | N | A | T | N | C |

SNPs or alleles

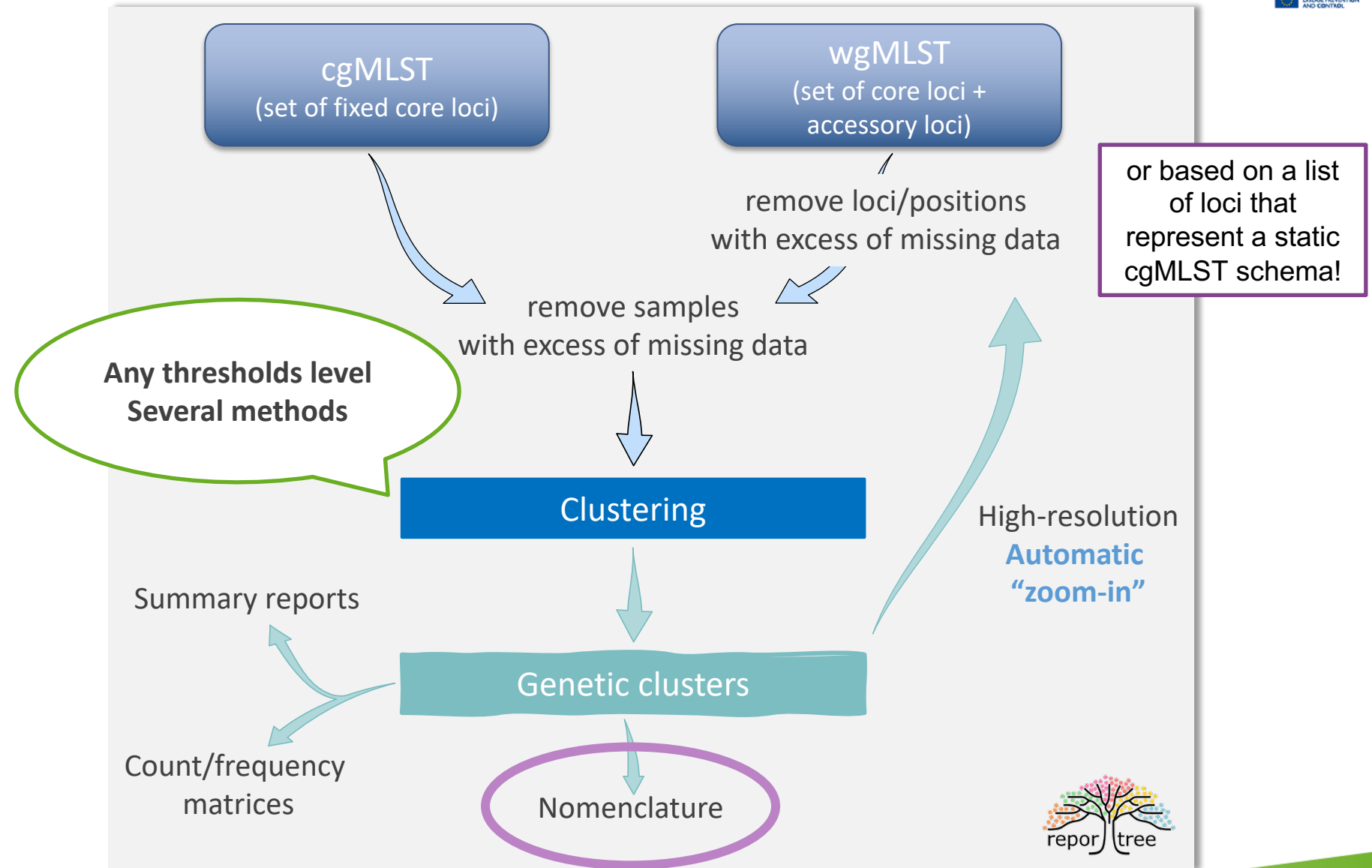


# ReportTree workflow

## Genetic data

|   |   |   |   |   |   |   |   |
|---|---|---|---|---|---|---|---|
| A | T | C | G | A | T | C | G |
| A | T | C | G | A | T | C | C |
| N | A | C | G | A | T | C | C |
| A | C | G | G | A | T | C | C |
| N | C | G | N | A | T | N | C |

SNPs or alleles



**How does it help me monitoring clusters  
and communicating?**

# Detect and track genetic clusters with ReporTree nomenclature system



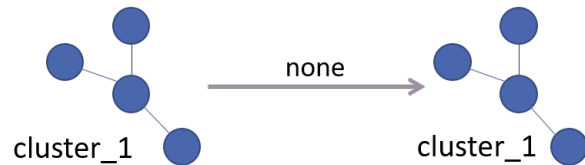
Cluster names are maintained in subsequent runs, facilitating the establishment of a stable local nomenclature...

# Detect and track genetic clusters with ReporTree nomenclature system

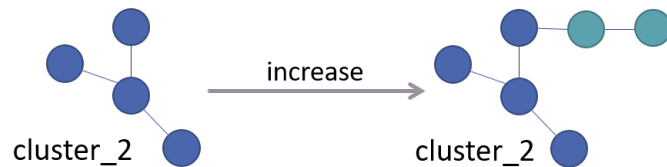
Cluster names are maintained in subsequent runs, facilitating the establishment of a stable local nomenclature...

## Most common cluster changes

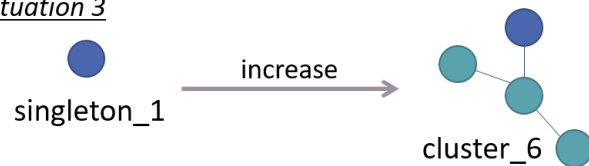
### Situation 1



### Situation 2



### Situation 3



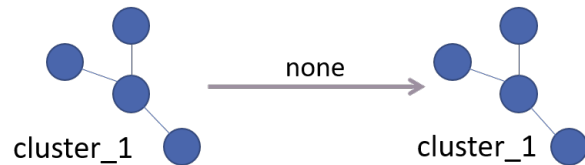
- sample used in previous run
- sample used in previous run
- new sample (i.e. not used in previous run)

# Detect and track genetic clusters with ReportTree nomenclature system

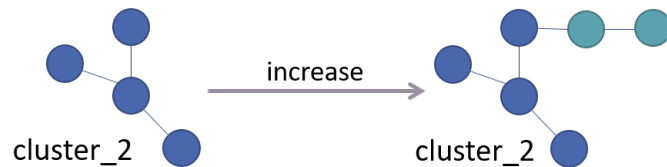
Cluster names are maintained in subsequent runs, facilitating the establishment of a stable local nomenclature...

## Most common cluster changes

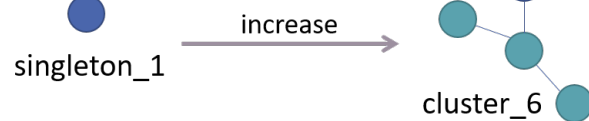
Situation 1



Situation 2

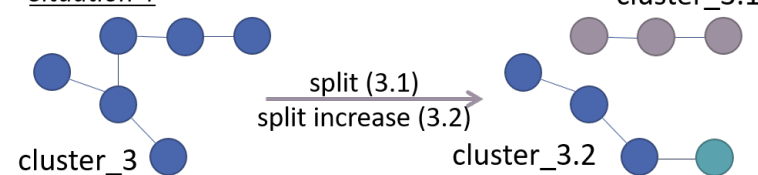


Situation 3

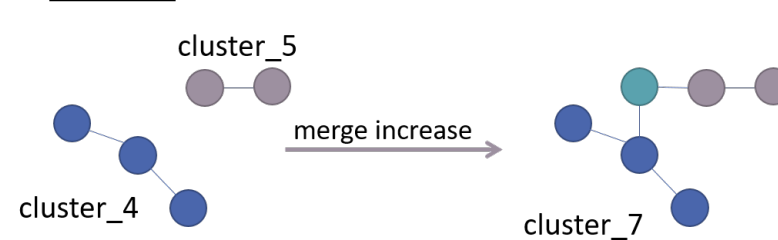


## Other cluster changes

Situation 4



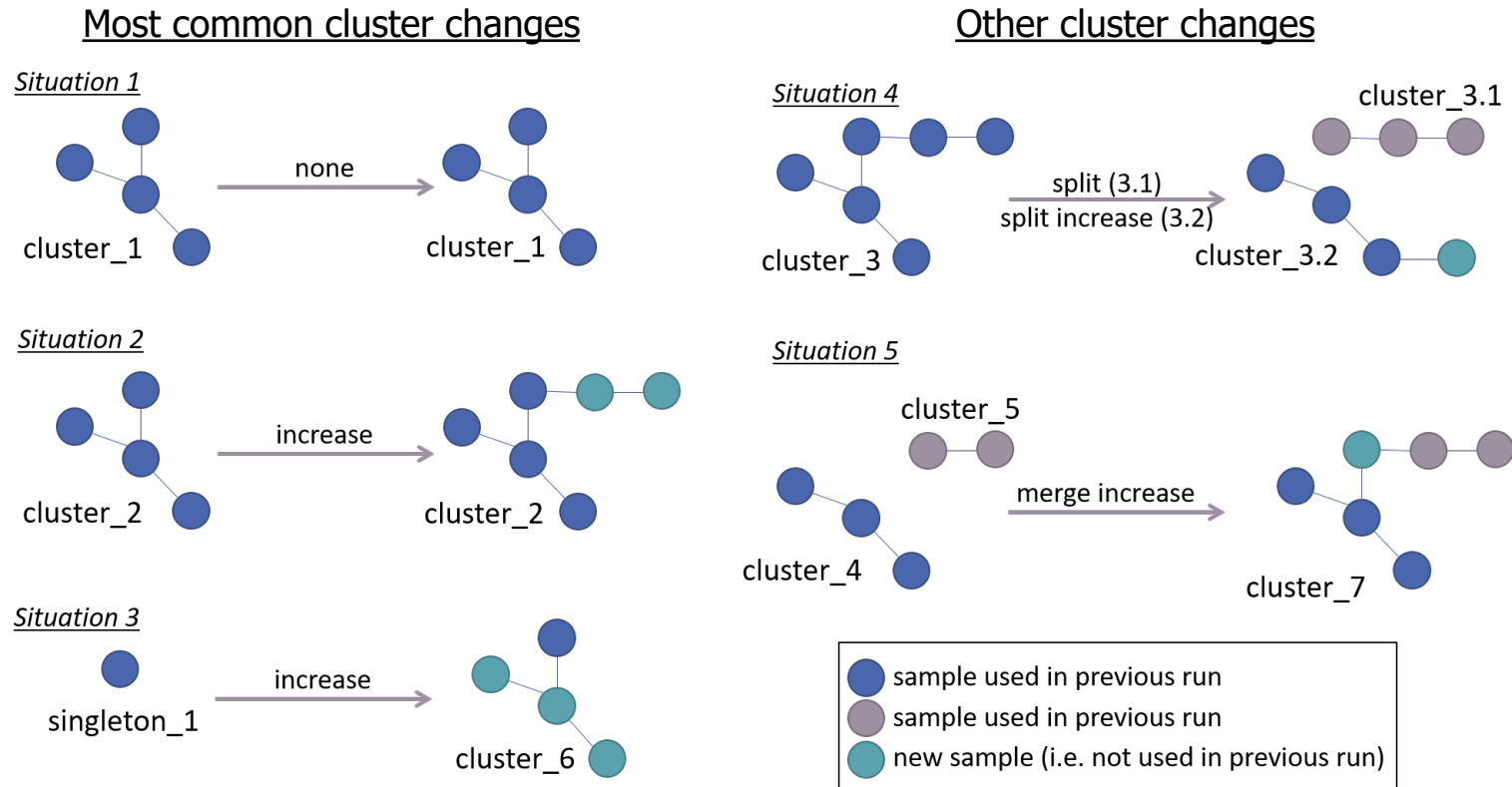
Situation 5



- sample used in previous run
- sample used in previous run
- new sample (i.e. not used in previous run)

# Detect and track genetic clusters with ReportTree nomenclature system

Cluster names are maintained in subsequent runs, facilitating the establishment of a stable local nomenclature...



**Cluster names  
always reflect the  
structure of the tree**



# Facilitating monitoring of main circulating lineages and ongoing outbreaks

Cluster names are maintained in subsequent runs, facilitating the establishment of a stable local nomenclature...

| partition  | cluster   | nomenclature_change | n_increase | cluster_length | first_seq_date | last_seq_date | source                                 | samples_increase                       |
|------------|-----------|---------------------|------------|----------------|----------------|---------------|----------------------------------------|----------------------------------------|
| MST-10x1.0 | cluster_1 | kept (none)         | 0          | 50             | 14/01/2023     | 02/06/2025    | Clinical (96.0%), Food (4.0%) (n = 50) |                                        |
| MST-10x1.0 | cluster_2 | kept (increase)     | 1          | 21             | 07/01/2024     | 01/09/2025    | Clinical (95.2%), Food (4.8%) (n =21)  | Sample_A                               |
| MST-10x1.0 | cluster_3 | kept (increase)     | 2          | 20             | 21/02/2025     | 01/09/2025    | Clinical (100.0%) (n = 20)             | Sample_B, Sample_C                     |
| MST-10x1.0 | cluster_4 | kept (none)         | 0          | 10             | 07/09/2024     | 24/06/2025    | Clinical (100.0%) (n = 10)             |                                        |
| MST-10x1.0 | cluster_5 | new                 | 4          | 4              | 09/08/2025     | 01/09/2025    | Clinical (100.0%) (n = 4)              | Sample_D, Sample_E, Sample_F, Sample_G |

# Facilitating monitoring of main circulating lineages and ongoing outbreaks

Cluster names are maintained in subsequent runs, facilitating the establishment of a stable local nomenclature...

| partition  | cluster   | nomenclature_change | n_increase | cluster_length | first_seq_date | last_seq_date | source                                 | samples_increase                       |
|------------|-----------|---------------------|------------|----------------|----------------|---------------|----------------------------------------|----------------------------------------|
| MST-10x1.0 | cluster_1 | kept (none)         | 0          | 50             | 14/01/2023     | 02/06/2025    | Clinical (96.0%), Food (4.0%) (n = 50) |                                        |
| MST-10x1.0 | cluster_2 | kept (increase)     | 1          | 21             | 07/01/2024     | 01/09/2025    | Clinical (95.2%), Food (4.8%) (n =21)  | Sample_A                               |
| MST-10x1.0 | cluster_3 | kept (increase)     | 2          | 20             | 21/02/2025     | 01/09/2025    | Clinical (100.0%) (n = 20)             | Sample_B, Sample_C                     |
| MST-10x1.0 | cluster_4 | kept (none)         | 0          | 10             | 07/09/2024     | 24/06/2025    | Clinical (100.0%) (n = 10)             |                                        |
| MST-10x1.0 | cluster_5 | new                 | 4          | 4              | 09/08/2025     | 01/09/2025    | Clinical (100.0%) (n = 4)              | Sample_D, Sample_E, Sample_F, Sample_G |

# Facilitating monitoring of main circulating lineages and ongoing outbreaks

Cluster names are maintained in subsequent runs, facilitating the establishment of a stable local nomenclature...

| partition  | cluster   | nomenclature_change | n_increase | cluster_length | first_seq_date | last_seq_date | source                                 | samples_increase                       |
|------------|-----------|---------------------|------------|----------------|----------------|---------------|----------------------------------------|----------------------------------------|
| MST-10x1.0 | cluster_1 | kept (none)         | 0          | 50             | 14/01/2023     | 02/06/2025    | Clinical (96.0%), Food (4.0%) (n = 50) |                                        |
| MST-10x1.0 | cluster_2 | kept (increase)     | 1          | 21             | 07/01/2024     | 01/09/2025    | Clinical (95.2%), Food (4.8%) (n =21)  | Sample_A                               |
| MST-10x1.0 | cluster_3 | kept (increase)     | 2          | 20             | 21/02/2025     | 01/09/2025    | Clinical (100.0%) (n = 20)             | Sample_B, Sample_C                     |
| MST-10x1.0 | cluster_4 | kept (none)         | 0          | 10             | 07/09/2024     | 24/06/2025    | Clinical (100.0%) (n = 10)             |                                        |
| MST-10x1.0 | cluster_5 | new                 | 4          | 4              | 09/08/2025     | 01/09/2025    | Clinical (100.0%) (n = 4)              | Sample_D, Sample_E, Sample_F, Sample_G |

**All changes in the tree are recorded in a comprehensive report!**

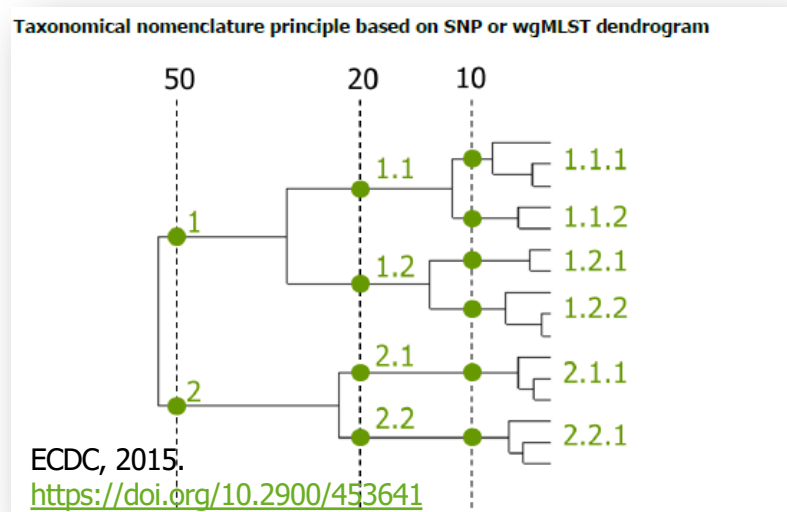
# Facilitating monitoring of main circulating lineages and ongoing outbreaks



Cluster names are maintained in subsequent runs, facilitating the establishment of a stable local nomenclature... and, if requested, a hierarchical nomenclature similar to “SNP-address”.

# Facilitating monitoring of main circulating lineages and ongoing outbreaks

Cluster names are maintained in subsequent runs, facilitating the establishment of a stable local nomenclature... and, if requested, a hierarchical nomenclature similar to “SNP-address”.

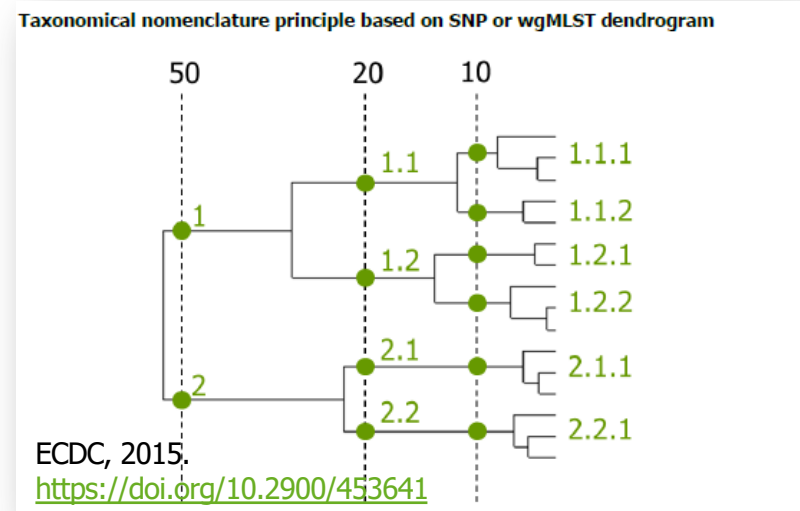


# Facilitating monitoring of main circulating lineages and ongoing outbreaks



Cluster names are maintained in subsequent runs, facilitating the establishment of a stable local nomenclature... and, if requested, a hierarchical nomenclature similar to “SNP-address”.

The user can set the hierarchical thresholds for nomenclature (e.g. 50,20,7,0)



| sample    | 50   | 20   | 7     | 0     | nomenclature_code_2025-10-01 |
|-----------|------|------|-------|-------|------------------------------|
| sample_01 | C564 | C550 | C647  | C38   | C564-C550-C647-C38           |
| sample_02 | C564 | C550 | C647  | C38   | C564-C550-C647-C38           |
| sample_03 | C564 | C550 | C647  | S1337 | C564-C550-C647-S1337         |
| sample_04 | S136 | S492 | S1397 | S4880 | S136-S492-S1397-S4880        |
| sample_05 | C452 | S28  | S927  | S4408 | C452-S28-S927-S4408          |
| sample_06 | C286 | C372 | C626  | S2861 | C286-C372-C626-S2861         |
| sample_07 | C286 | C372 | C626  | S2862 | C286-C372-C626-S2862         |
| sample_08 | C21  | C37  | C75   | C570  | C21-C37-C75-C570             |
| sample_09 | C21  | C37  | C75   | C570  | C21-C37-C75-C570             |

C – cluster  
S - singleton

# **How can I determine the nomenclature code levels?**

# Clustering stability regions

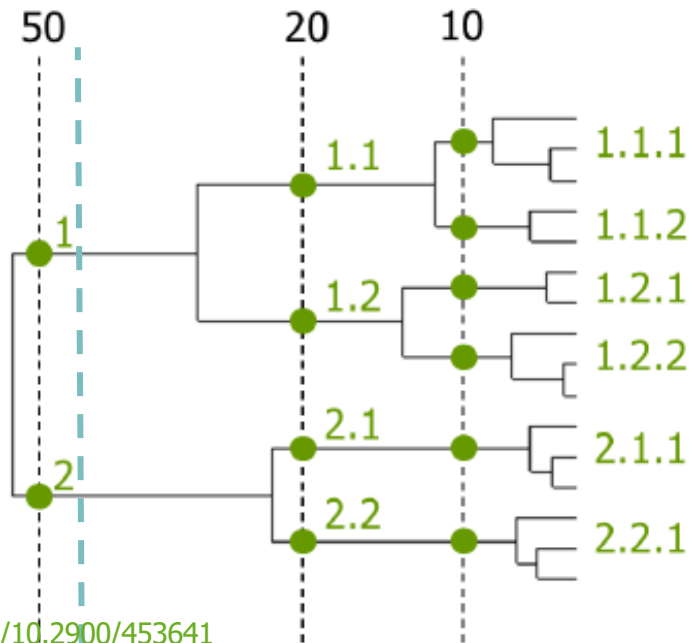
*Stability regions* are ranges of subsequent distance thresholds providing similar clustering results.



# Clustering stability regions

*Stability regions* are ranges of subsequent distance thresholds providing similar clustering results.

Taxonomical nomenclature principle based on SNP or wgMLST dendrogram



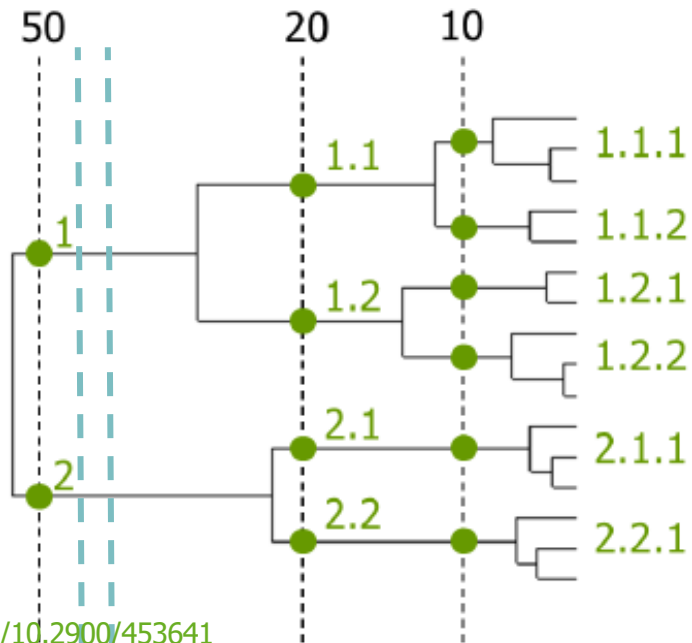
ECDC, 2015.

<https://doi.org/10.2900/453641>

# Clustering stability regions

*Stability regions* are ranges of subsequent distance thresholds providing similar clustering results.

Taxonomical nomenclature principle based on SNP or wgMLST dendrogram



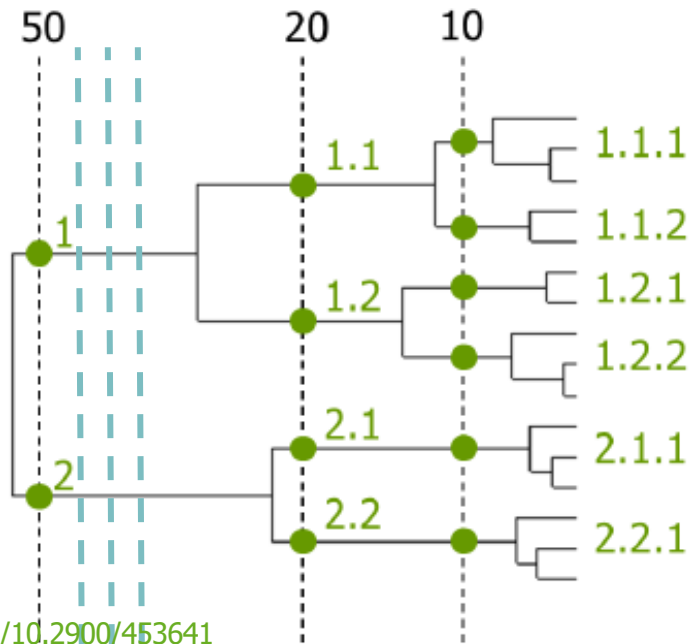
ECDC, 2015.

<https://doi.org/10.2900/453641>

# Clustering stability regions

*Stability regions* are ranges of subsequent distance thresholds providing similar clustering results.

Taxonomical nomenclature principle based on SNP or wgMLST dendrogram



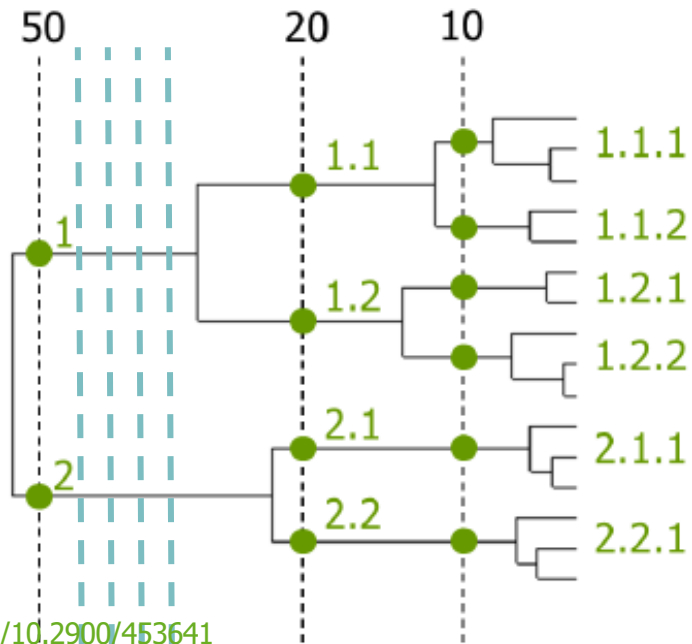
ECDC, 2015.

<https://doi.org/10.2900/453641>

# Clustering stability regions

*Stability regions* are ranges of subsequent distance thresholds providing similar clustering results.

Taxonomical nomenclature principle based on SNP or wgMLST dendrogram



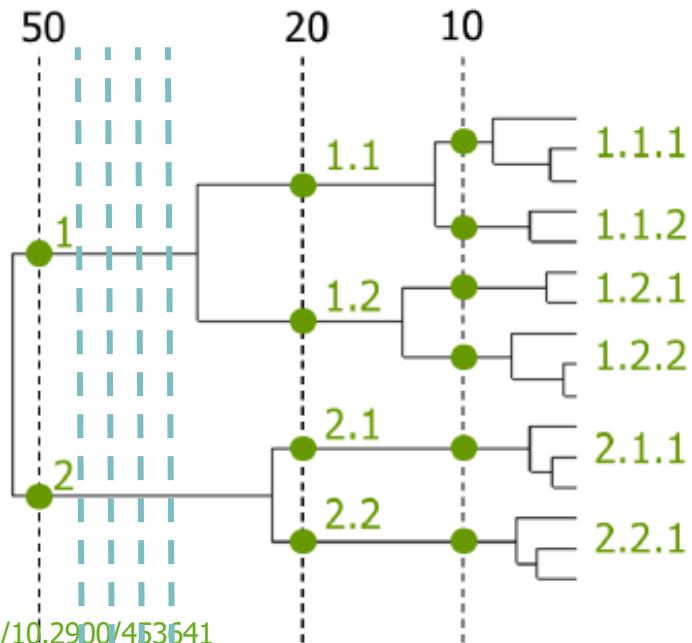
ECDC, 2015.

<https://doi.org/10.2900/453641>

# Clustering stability regions

*Stability regions* are ranges of subsequent distance thresholds providing similar clustering results.

Taxonomical nomenclature principle based on SNP or wgMLST dendrogram



ECDC, 2015.

<https://doi.org/10.2900/453641>

Candidate regions to define cutoffs for pathogen-specific **nomenclature design**, as they are less likely to be “disrupted” by the introduction of new samples

# Identification of candidate thresholds for nomenclature design



Neighborhood Adjusted Wallace coefficient (nAWC) – evaluates the agreement between two clustering methods (e.g. subsequent distance thresholds)

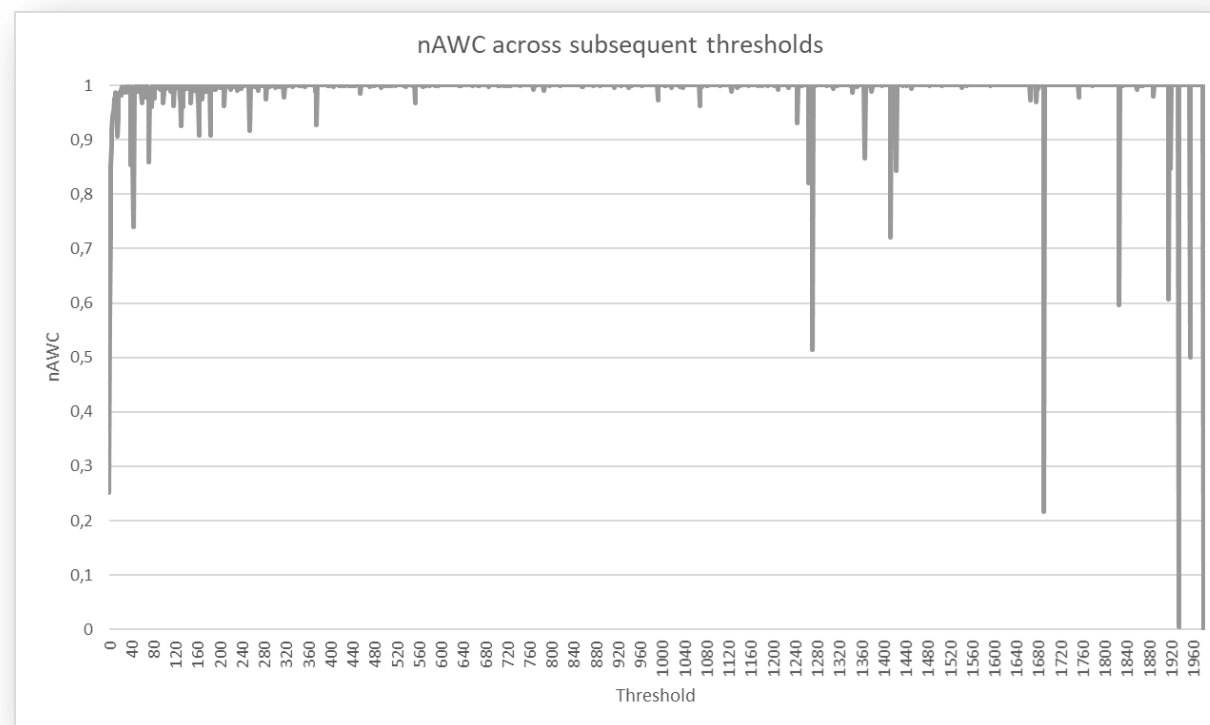
# Identification of candidate thresholds for nomenclature design

Neighborhood Adjusted Wallace coefficient (nAWC) – evaluates the agreement between two clustering methods (e.g. subsequent distance thresholds)

*comparing\_partitions\_v2.py*

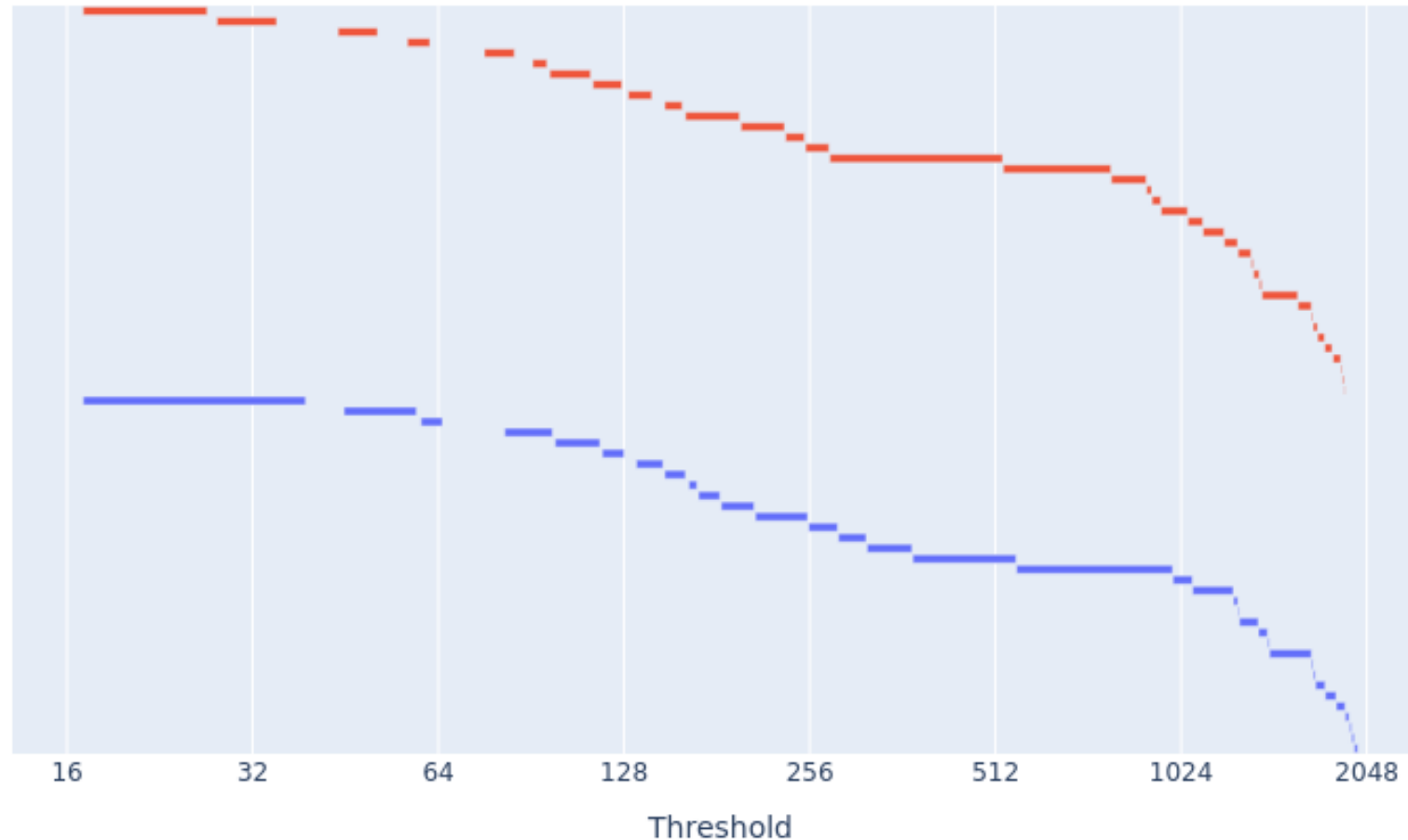


Determines the nAWC at  
consecutive thresholds  
("n + 1" → "n")



# Translating nAWC into practice

Threshold ranges in which clustering remains similar\*  
- cluster stability regions -

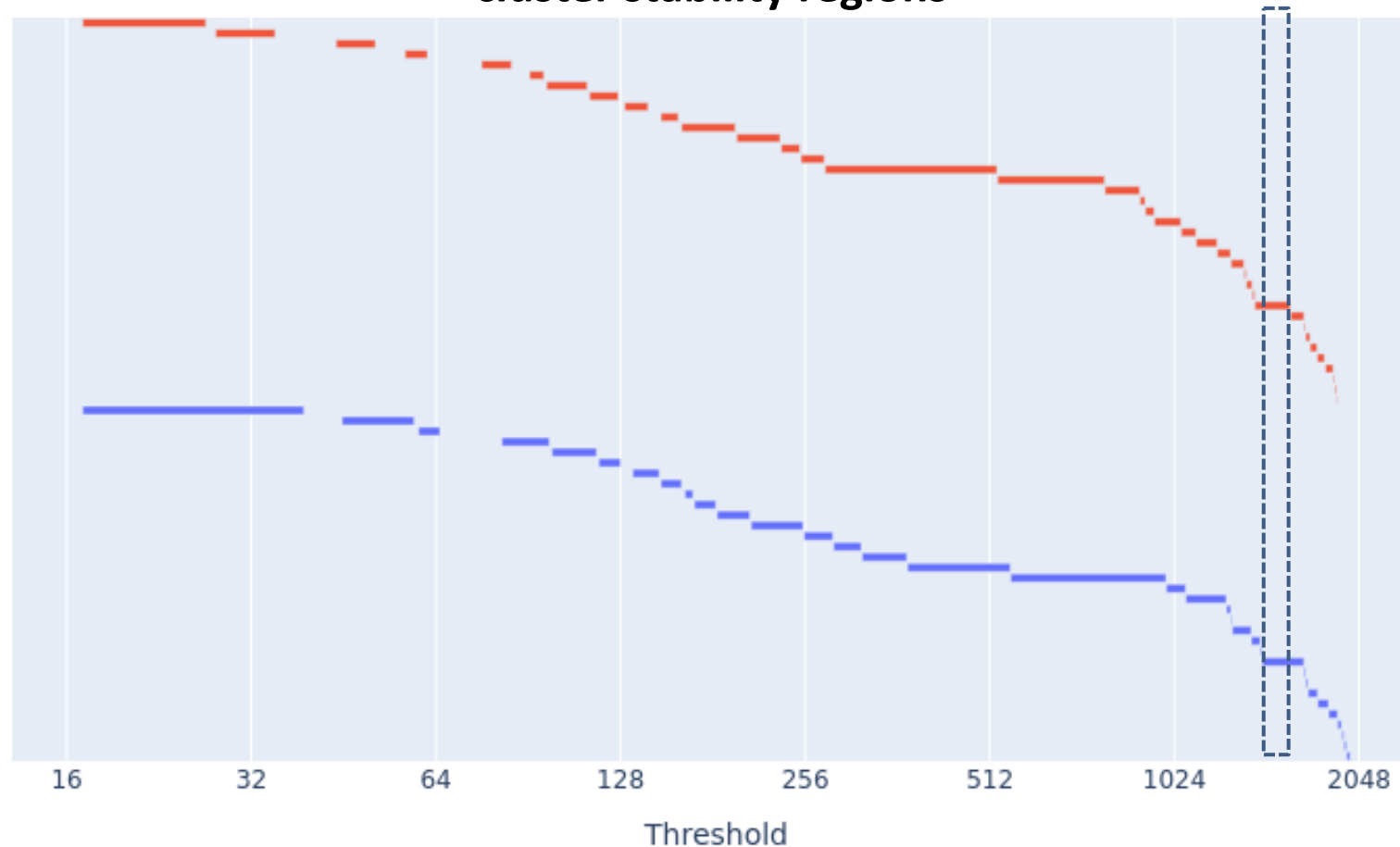


*\*Blocks identified with comparing\_partitions\_v2.py for two pipelines running the same dataset  
(visualization: EvalTree module of ACDTree toolkit)*



# Translating nAWC into practice

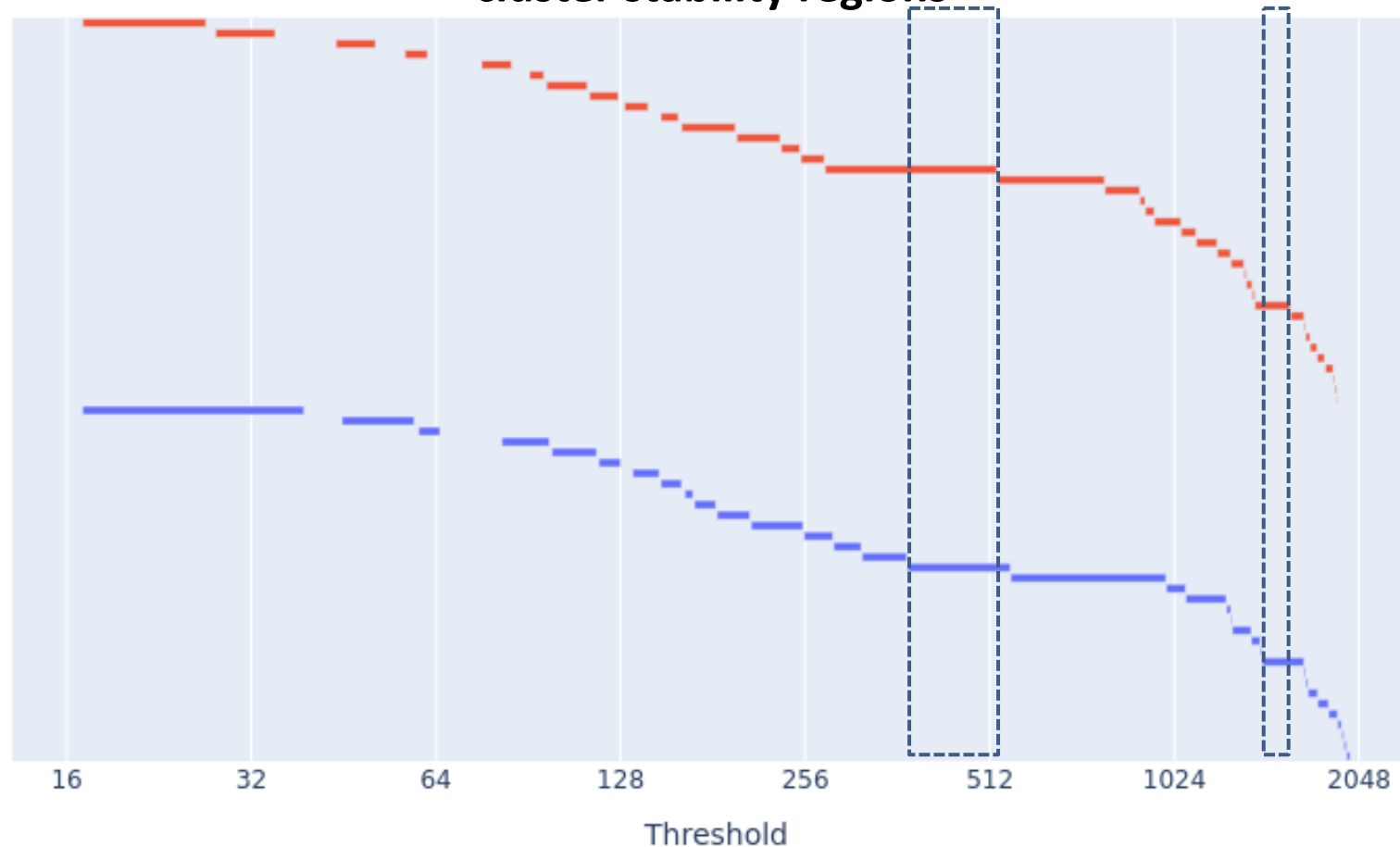
Threshold ranges in which clustering remains similar\*  
- cluster stability regions -



\*Blocks identified with `comparing_partitions_v2.py` for two pipelines running the same dataset  
(visualization: `EvalTree` module of `ACDTree` toolkit)

# Translating nAWC into practice

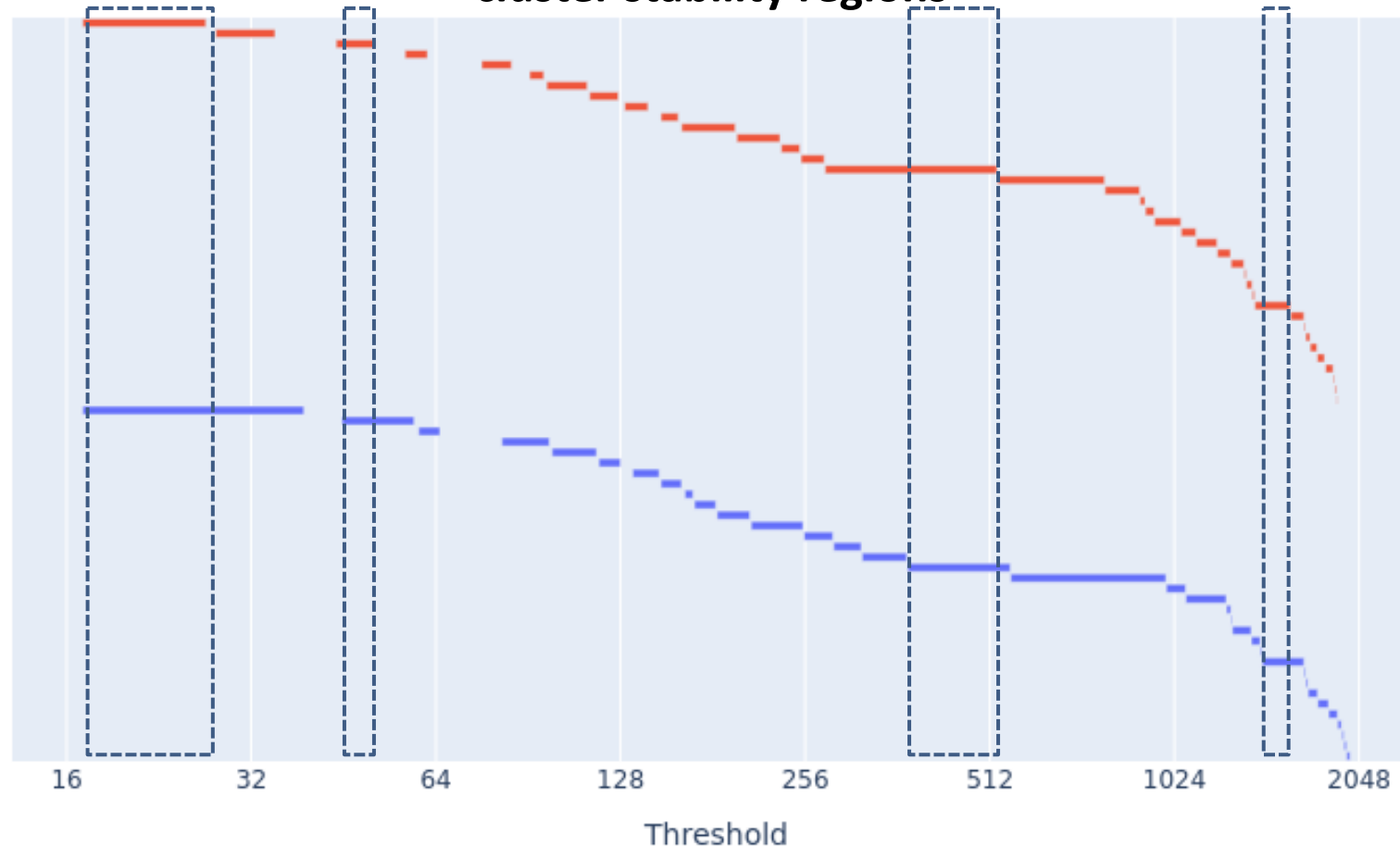
Threshold ranges in which clustering remains similar\*  
- cluster stability regions -



\*Blocks identified with `comparing_partitions_v2.py` for two pipelines running the same dataset  
(visualization: `EvalTree` module of `ACDTree` toolkit)

# Translating nAWC into practice

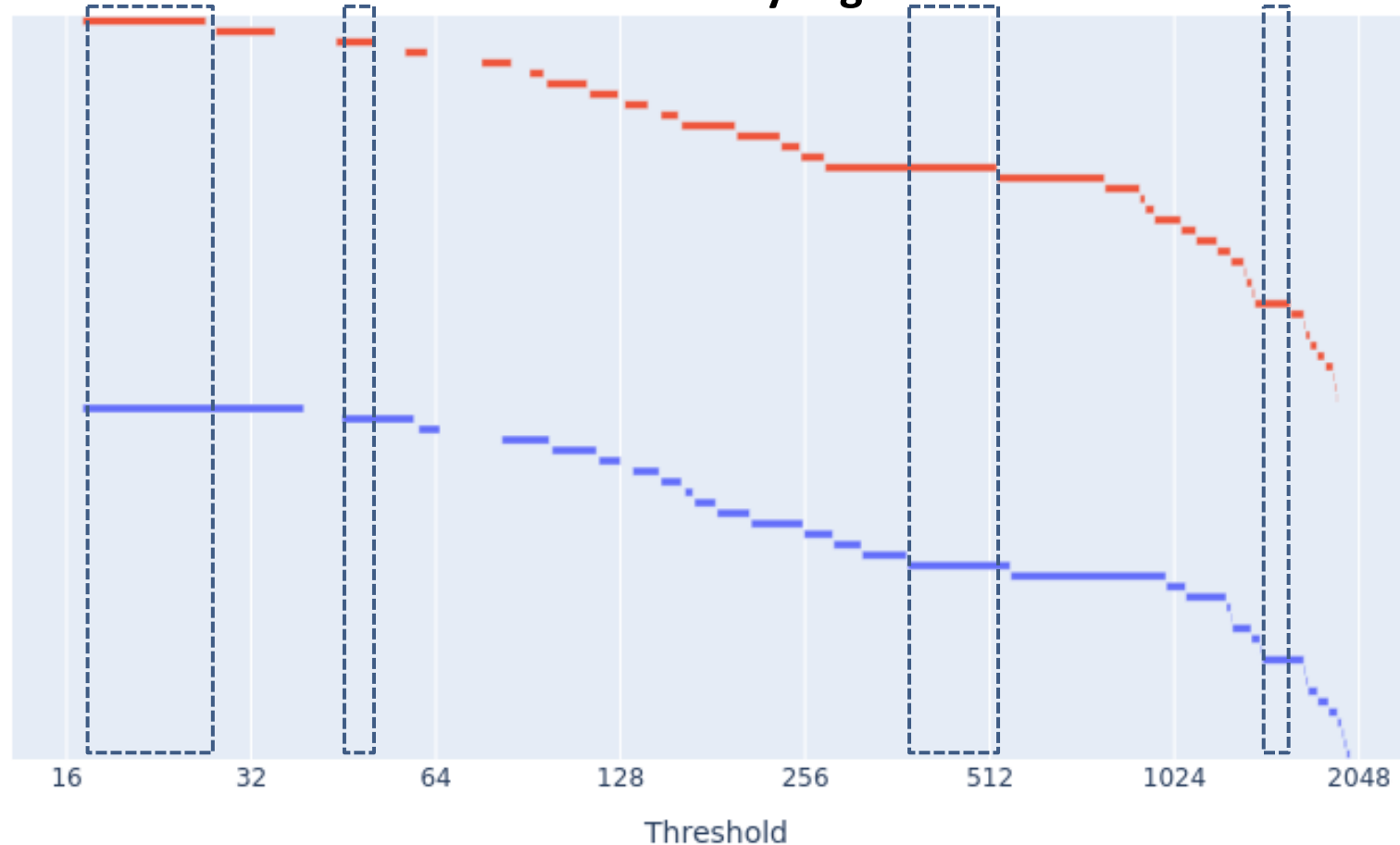
Threshold ranges in which clustering remains similar\*  
- cluster stability regions -



\*Blocks identified with `comparing_partitions_v2.py` for two pipelines running the same dataset  
(visualization: `EvalTree` module of `ACDTree` toolkit)

# Translating nAWC into practice

Threshold ranges in which clustering remains similar\*  
- cluster stability regions -



--nomenclature-code-levels  
1500,500,50,20

\*Blocks identified with `comparing_partitions_v2.py` for two pipelines running the same dataset  
(visualization: `EvalTree` module of `ACDTree` toolkit)

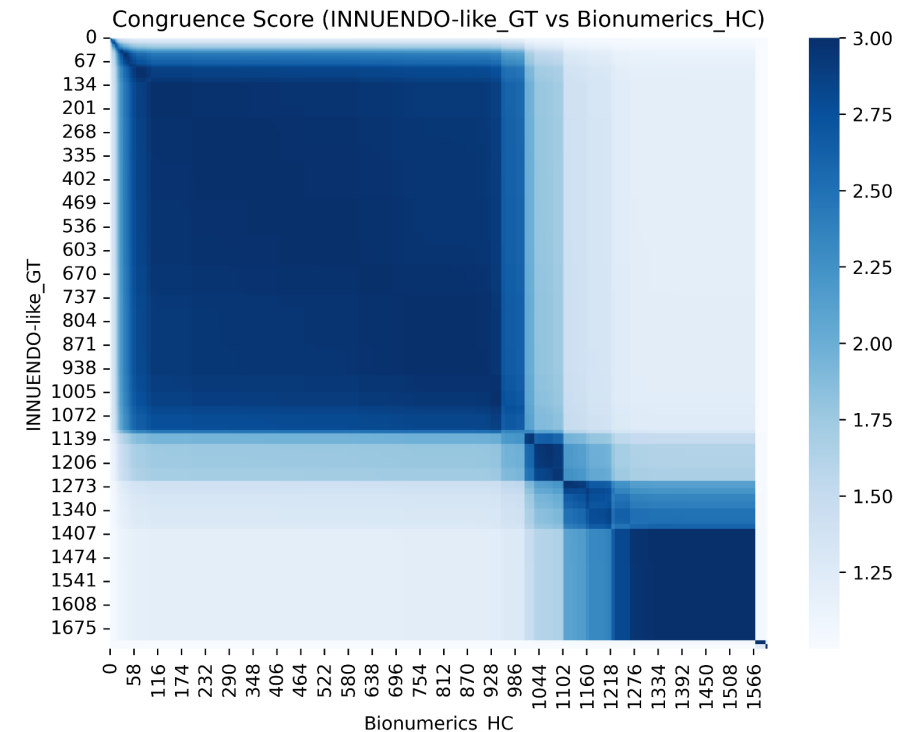
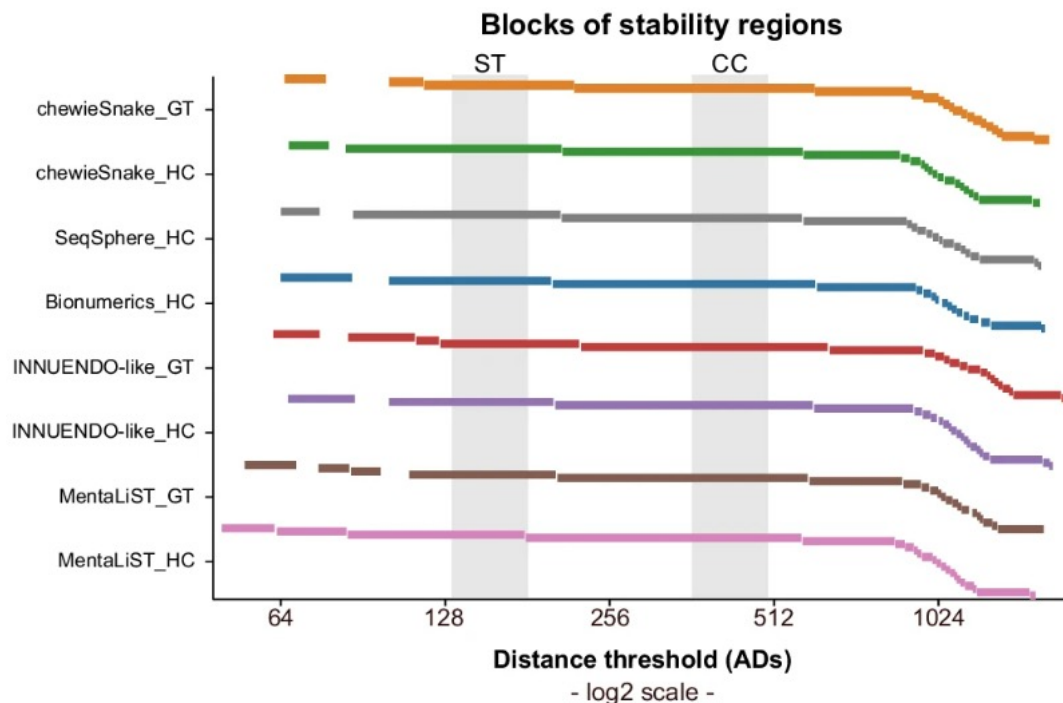
# Cluster stability regions across pipelines and WGS backward compatibility with other typing solutions



- As they reflect the population structure, some cluster stability regions may correspond to some genetic-based traditional typing -> this can also be a criteria to use it in the nomenclature code

# Cluster stability regions across pipelines and WGS backward compatibility with other typing solutions

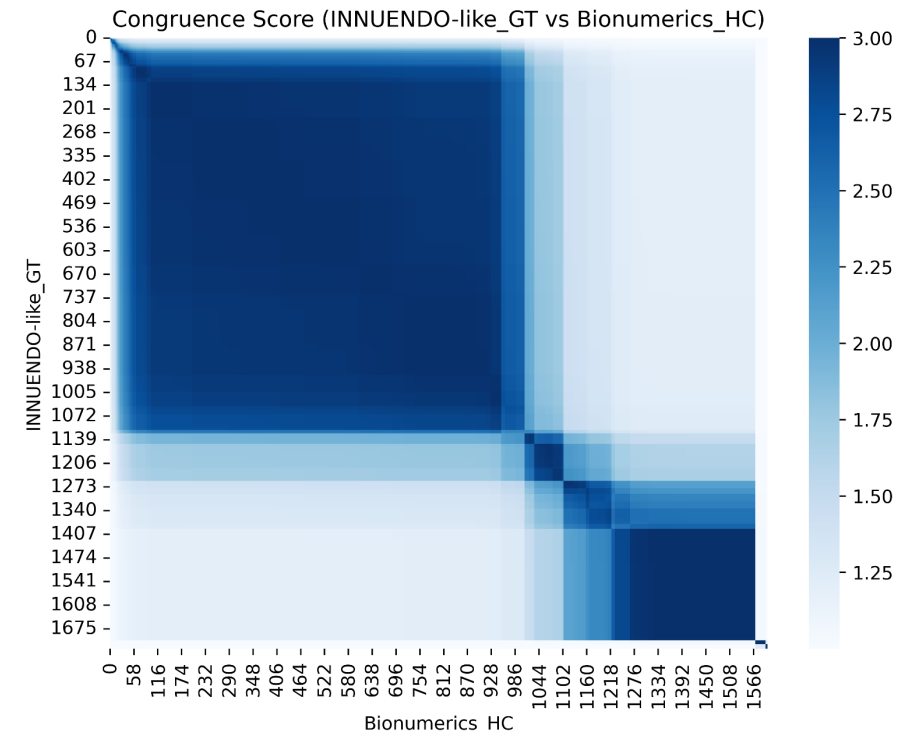
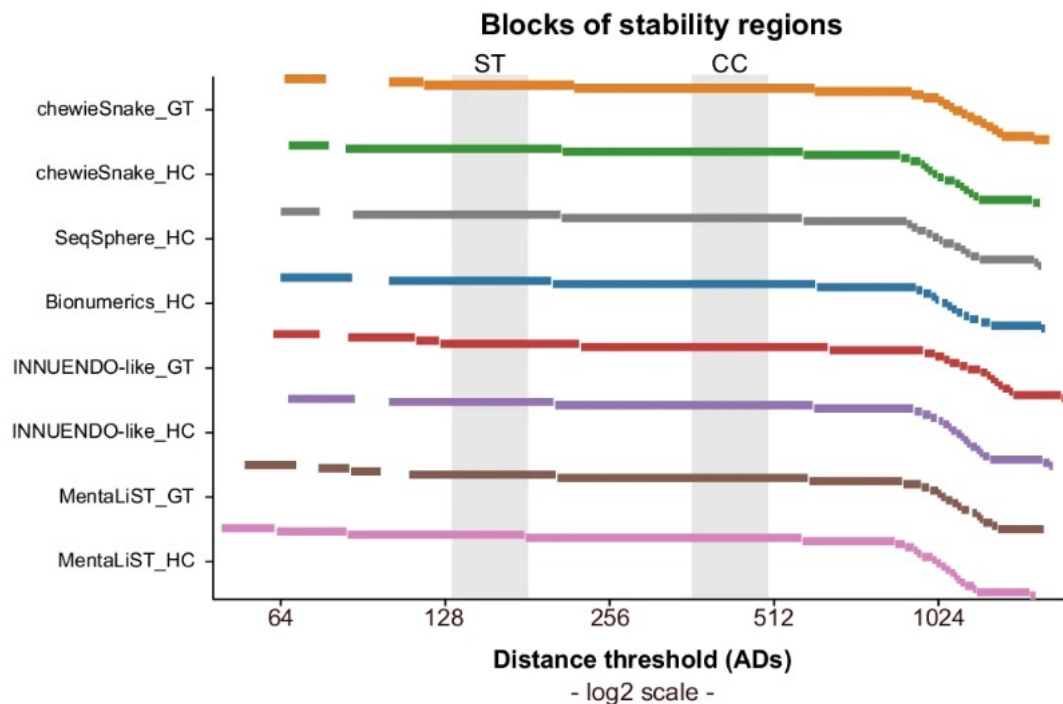
- As they reflect the population structure, some cluster stability regions may correspond to some genetic-based traditional typing -> this can also be a criteria to use it in the nomenclature code



*E.g. large-scale multicountry inter-pipeline cluster congruence  
(Food- and water-borne bacterial pathogens)*

# Cluster stability regions across pipelines and WGS backward compatibility with other typing solutions

- As they reflect the population structure, some cluster stability regions may correspond to some genetic-based traditional typing -> this can also be a criteria to use it in the nomenclature code



**Given its input flexibility and by providing clustering information at all possible threshold levels, ReporTree facilitates large-scale congruence analyses!**

# Exercise – genomic surveillance of *Salmonella enterica*



- (Pre-webinar: install ReporTree)
- Download the provided material (reportree\_exercise.zip) and paste it in ReporTree folder
- `unzip reportree_exercise.zip`
- Check its contents (`ls reportree_exercise/`):
  - `run_last_week/` – folder with the results of a previous ReporTree run
  - `run_this_week/` – folder with the input files necessary for this ReporTree run:
    - `Se_beone_alleles.tsv` – wgMLST allele matrix for all samples in the dataset (including samples added since “last week”)
    - `Se_beone_metadata.tsv` – metadata table for all samples in the dataset
    - `cgMLST_Salmonella_enterica_chewie-NS.txt` – List of cgMLST loci



# Exercise – genomic surveillance of *Salmonella enterica*



## Genomic surveillance command line:

```
python /PATH/TO/reporttree.py -a Se_beone_alleles.tsv \  
    -m Se_beone_metadata.tsv \  
    -out Se_beone_GT \  
    -l cgMLST_Salmonella_enterica_chewie-NS.txt \  
    --loci-called 0.95 \  
    -thr 1,2,5,10,25,100,500,1000 \  
    --analysis grapetree \  
    --nomenclature-file ../run_last_week/Se_beone_last_week_GT_partitions.tsv \  
    --nomenclature-code 500,100,25,10 \  
    --columns_summary_report MLST_ST,Predicted_serotype,Country,n_Country,Source,  
n_Source,first_seq_date,last_seq_date,timespan_days
```

# Exercise – genomic surveillance of *Salmonella enterica*



## Genomic surveillance command line:

```
python /PATH/TO/reporttree.py -a Se_beone_alleles.tsv \  
    -m Se_beone_metadata.tsv \  
    -out Se_beone_GT \  
    -l cgMLST_Salmonella_enterica_chewie-NS.txt \  
    --loci-called 0.95 \  
    -thr 1,2,5,10,25,100,500,1000 \  
    --analysis grapetree \  
    --nomenclature-file ../run_last_week/Se_beone_last_week_GT_partitions.tsv \  
    --nomenclature-code 500,100,25,10 \  
    --columns_summary_report MLST_ST,Predicted_serotype,Country,n_Country,Source,  
n_Source,first_seq_date,last_seq_date,timespan_days
```

+ Zoom-in analysis: `--sample_of_interest ERR10442916 --zoom-cluster-of-interest 5 --site-inclusion 1.0 --unzip`

+ Finding cluster stability regions: `--partitions2report stability_regions` ~~`--thr 1,2,5,10,25,100,500,1000`~~

# In summary

List of learning points in this session:

# In summary

List of learning points in this session:

- ReporTree is a flexible solution that automatically identifies genetic clusters at any (or all) distance thresholds and generates surveillance-oriented reports

# In summary

List of learning points in this session:

- ReporTree is a flexible solution that automatically identifies genetic clusters at any (or all) distance thresholds and generates surveillance-oriented reports
- It can be smoothly implemented in **routine surveillance**, providing a cluster nomenclature system that facilitates monitoring main circulating lineages and ongoing outbreaks

# In summary

List of learning points in this session:

- ReporTree is a [flexible solution that automatically identifies genetic clusters](#) at any (or all) distance thresholds and generates surveillance-oriented reports
- It can be smoothly implemented in **routine surveillance**, providing a [cluster nomenclature system](#) that facilitates monitoring main circulating lineages and ongoing outbreaks
- When dealing with an wgMLST schema, ReporTree automatically performs a [dynamic wgMLST approach](#) to zoom-in in any cgMLST cluster of interest

# In summary

List of learning points in this session:

- ReporTree is a flexible solution that automatically identifies genetic clusters at any (or all) distance thresholds and generates surveillance-oriented reports
- It can be smoothly implemented in **routine surveillance**, providing a cluster nomenclature system that facilitates monitoring main circulating lineages and ongoing outbreaks
- When dealing with an wgMLST schema, ReporTree automatically performs a dynamic wgMLST approach to zoom-in in any cgMLST cluster of interest
- As a complement, it can identify clustering stability regions, a useful approach for pathogen specific **nomenclature design**

# In summary

List of learning points in this session:

- ReporTree is a flexible solution that automatically identifies genetic clusters at any (or all) distance thresholds and generates surveillance-oriented reports
- It can be smoothly implemented in **routine surveillance**, providing a cluster nomenclature system that facilitates monitoring main circulating lineages and ongoing outbreaks
- When dealing with an wgMLST schema, ReporTree automatically performs a dynamic wgMLST approach to zoom-in in any cgMLST cluster of interest
- As a complement, it can identify clustering stability regions, a useful approach for pathogen specific **nomenclature design**

**ReporTree contributes to a sustainable and efficient public health  
genomics-informed pathogen surveillance**



# Further reading

Stay tuned!  
Updates coming soon!!!

ReporTree GitHub:

<https://github.com/insapathogenomics/ReporTree>

Clustering stability regions:

<https://github.com/insapathogenomics/ComparingPartitions>

<https://github.com/insapathogenomics/ACDTree>



# References

- Dallman et al. (2018) SnapperDB: a database solution for routine sequencing analysis of bacterial isolates. *Bioinformatics*. 34(17):3028-3029. doi: 10.1093/bioinformatics/bty212
- ECDC (2025) One Health (<https://www.ecdc.europa.eu/en/one-health>)
- Francisco et al. (2012) PHYLOViZ: phylogenetic inference and data visualization for sequence based typing methods. *BMC Bioinformatics*. 13:87. doi: 10.1186/1471-2105-13-87
- Mixão et al. (2023) ReporTree: a surveillance-oriented tool to strengthen the linkage between pathogen genetic clusters and epidemiological data. *Genome Med*. 15(1):43. doi: 10.1186/s13073-023-01196-1
- Mixão et al. (2025) Multi-country and intersectoral assessment of cluster congruence between pipelines for genomics surveillance of foodborne pathogens. *Nat Commun*. 16(1):3961. doi: 10.1038/s41467-025-59246-8
- Severiano et al. (2011) Adjusted Wallace coefficient as a measure of congruence between typing methods. *J Clin Microbiol*. 49(11):3997-4000. doi: 10.1128/JCM.00624-11
- WHO (2022) WHO global genomic surveillance strategy for pathogens with pandemic and epidemic potential 2022-2032 (<https://www.who.int/initiatives/genomic-surveillance-strategy>)
- Zhou et al. (2018) GrapeTree: visualization of core genomic relationships among 100,000 bacterial pathogens. *Genome Res*. 28(9):1395-1404. doi: 10.1101/gr.232397.117

**Genomics and Bioinformatics Unit  
of the National Institute of Health  
Doutor Ricardo Jorge (INSA)**

Daniel Sobral  
Miguel Pinto  
João Paulo Gomes  
Vítor Borges



[veronica.mixao@insa.min-saude.pt](mailto:veronica.mixao@insa.min-saude.pt)

# Acknowledgements

The creation of this training material was commissioned by ECDC to National Institute of Health Doutor Ricardo Jorge (INSA) with the direct involvement of Verónica Mixão.

ReporTree development was supported by co-funding from the European Union's Horizon 2020 Research and Innovation program under grant agreement No 773830: One Health European Joint Programme.

ReporTree expansion and continuous maintenance is supported by national funds through FCT - Foundation for Science and Technology, I.P., in the frame of Individual CEEC 2022.00851.CEECIND/CP1748/CT0001, by the European Union project "Sustainable use and integration of enhanced infrastructure into routine genome-based surveillance and outbreak investigation activities in Portugal" - GENE0 [101113460] on behalf of the EU4H programme [EU4H-2022-DGA-MS-IBA-01-02], and by the DURABLE "Research Network against Epidemics" project. DURABLE is co-funded by The European Commission Union under the EU4Health Programme (EU4H) [101102733]. However, views and opinions expressed are those of the authors only and do not necessarily reflect those of the European Union or the European Health and Digital Executive Agency. Neither the European Union nor the granting authority can be held responsible.