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

Verónica Mixão[1], Miguel Pinto[1], Daniel Sobral[1], Adriano Di Pasquale[2], João Paulo Gomes[1], Vítor Borges[1]
veronica.mixao@insa.min-saude.pt
[1]Genomics and Bioinformatics Unit, Department of Infectious Diseases National Institute of Health Dr. Ricardo Jorge - Lisbon (Portugal)
[2]GENPAT, Istituto Zooprofilattico Sperimentale dell'Abruzzo e del Molise "Giuseppe Caporale" (IZSAM) – Teramo (Italy)

Motivation

Genomics-informed pathogen surveillance strengthens public health decision-making. A pivotal outcome of genomics surveillance is the identification of genetic clusters and their characterization in terms of geotemporal spread or linkage to clinical data. This task often consists of the visual exploration of (large) phylogenetic trees and associated metadata, being time-consuming and difficult to reproduce. As such, we aimed to create an **automated tool that facilitates and speeds-up the detection of genetic clusters and their linkage to epidemiological data.**

GitHub docker

SCAN ME

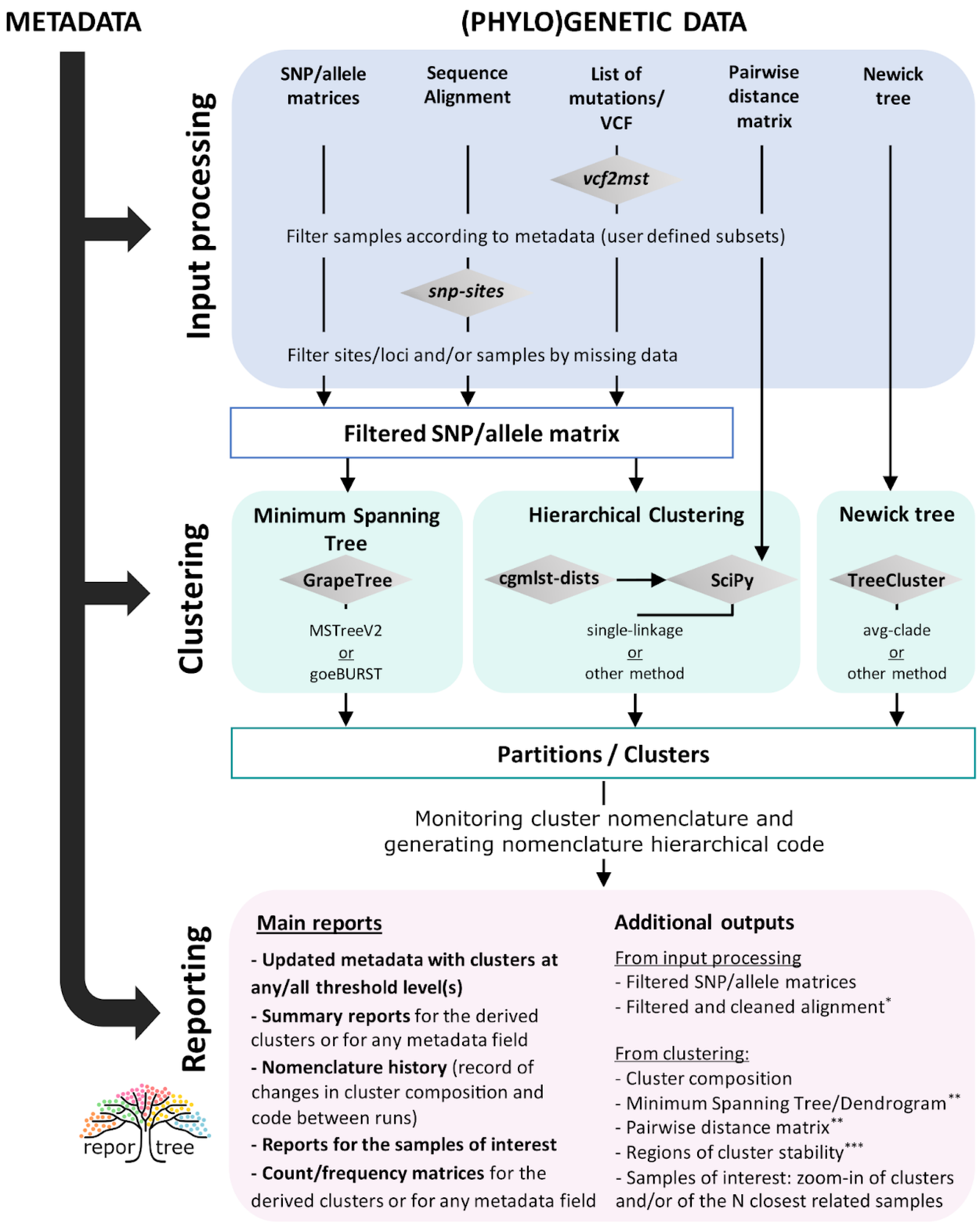
ReporTree can help you to…

- identify **genetic clusters at any threshold level(s)** of a tree, SNP or cg/wgMLST matrix, VCF files, sequence alignment, or distance matrix
- obtain **summary reports with the statistics/trends** (e.g., timespan, location, cluster/group composition, age distribution etc.) for the derived genetic clusters or for any other provided grouping variable (e.g., clade, lineage, ST, vaccination status, etc.)
- identify the phylogenetic context of **samples of interest** through an automated zoom-in on their clusters and/or through an automated in-depth analysis with the N closest related samples (particularly useful for wgMLST and alignment-based analyses)
- maintain **cluster nomenclature** between runs and generate **hierarchical codes** at your levels of interest
- identify regions of cluster stability (i.e., threshold ranges with similar cluster composition), a key step for pathogen-specific nomenclature design

Implementation

ReporTree is an **open-source** tool implemented in **python 3.8** that represents a **flexible** solution to obtain clustering information at any sample distance thresholds (partitions), either for species that require a cg/wgMLST analysis or for those that rely on SNPs/multiple sequence alignments for tree reconstruction.

ReporTree pipeline can be divided into three major steps:

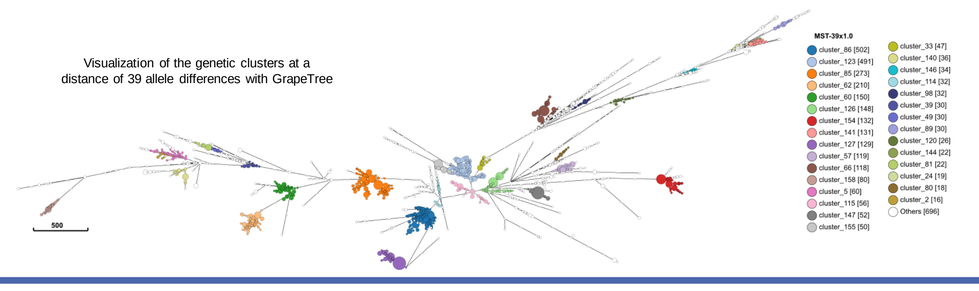


Validation

We reproduced the extensive genomics analysis of the bacterial pathogen ***Neisseria gonorrhoeae*** performed by Pinto et al., 2021. In this study, 3,791 *N. gonorrhoeae* genomes from isolates collected across Europe were analyzed with a cgMLST approach. With a single command line, we were able to obtain similar results in 2min 02s.

		Time		Geography		Traditional typing		Azithromycin		Penicillins
cluster	cluster_ length	first_seq_date	last_seq_date	country	n_ country	MLST	NG_MAST	23SrRNA_A2045G	23SrRNA_C2597T	blaTEM
cluster_86	502	06/01/2007	10/10/2017	United Kingdom (42.4%), Portugal (20.3%), Spain (4.6%),... (n = 502)	20	1901 (83.9%), 1579 (6.6%), 7360 (4.8%),...	1407 (55.8%), 2212 (3.4%), 3709 (2.6%),...	no (100.0%) (n = 502)	no (95.6%), het (2.4%), yes (2.0%) (n = 502)	no (94.0%), yes (6.0%) (n = 502)
cluster_123	491	09/01/2009	30/11/2017	United Kingdom (72.1%), Portugal (4.7%), Netherlands (3.9%),... (n = 491)	19	9363 (52.7%), 11428 (18.9%), 11463 (11.2%),...	2992 (58.5%), 3935 (8.8%), 7164 (4.5%),...	no (99.8%), yes (0.2%) (n = 491)	no (97.4%), yes (1.6%), het (1.0%) (n = 491)	no (96.3%), yes (3.7%) (n = 491)
cluster_85	273	09/01/2004	30/12/2014	United Kingdom (64.1%), Portugal (20.5%), Greece (3.7%),... (0.4%) (n = 273)	15	1901 (98.9%), 11992 (0.4%), ~1901 (0.4%),...	225 (67.4%), 5967 (3.7%), 19156 (3.3%),...	no (100.0%) (n = 273)	no (99.6%), yes (0.4%) (n = 273)	no (96.7%), yes (3.3%) (n = 273)
cluster_62	210	13/07/2010	20/12/2017	United Kingdom (53.3%), Portugal (11.0%), Netherlands (6.2%),... (n = 210)	17	7363 (97.6%), 1587 (1.4%), 11657 (0.5%),...	2400 (56.2%), 6360 (14.3%), 4943 (4.8%),...	no (100.0%) (n = 210)	no (94.3%), yes (3.3%), het (2.4%) (n = 210)	no (93.8%), yes (6.2%) (n = 210)
cluster_60	150	06/01/2004	21/11/2017	United Kingdom (40.7%), Portugal (20.7%), Latvia (12.0%),... (0.7%) (n = 150)	13	1579 (100.0%) (n = 150)	21 (32.0%), 1034 (19.3%), 5 (16.7%),...	no (100.0%) (n = 150)	no (99.3%), yes (0.7%) (n = 150)	no (97.3%), yes (2.7%) (n = 150)
cluster_126	148	06/01/2003	02/01/2017	United Kingdom (77.0%), Slovakia (6.1%), Ireland (6.1%),... (0.7%) (n = 148)	7	1580 (77.0%), 8126 (19.6%), 13739 (1.4%),...	9768 (38.5%), 359 (18.2%), 649 (12.2%),...	yes (56.8%), no (28.4%), het (14.9%) (n = 148)	no (95.9%), yes (3.4%), het (0.7%) (n = 148)	no (99.3%), yes (0.7%) (n = 148)

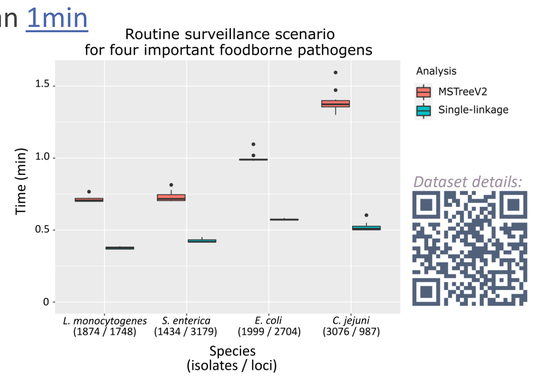
Visualization of the genetic clusters at a distance of 39 allele differences with GrapeTree



Benchmarking

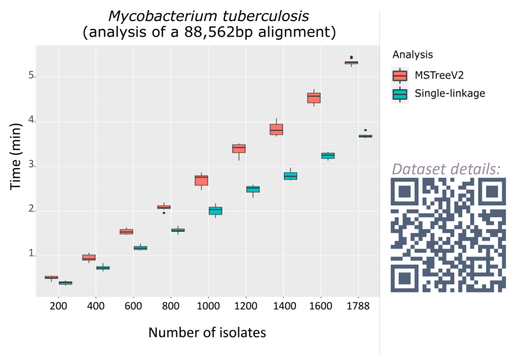
cg/wgMLST workflow

Using four diverse datasets of foodborne bacterial pathogens, ReporTree identifies genetic clusters at **potential outbreak level** and performs their characterization in less than 1min



alignment-based SNP workflow

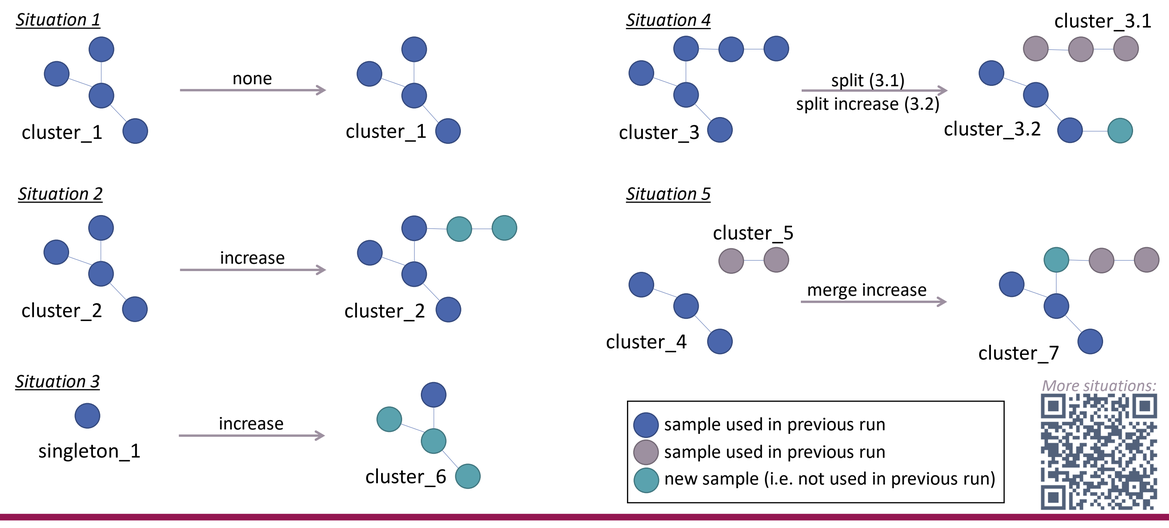
Using a diverse dataset of *Mycobacterium tuberculosis*, ReporTree identifies genetic clusters at **all possible levels** and performs their characterization in less than 6 min



ReporTree can be smoothly implemented in routine surveillance, with negligible computational and time costs

Cluster nomenclature (optional)

To facilitate routine surveillance and cluster monitoring over time, ReporTree can use the information of the partitions table of a previous run to (re)name the clusters in the current run. Below, we show a summary of the behavior of the **"Cluster Nomenclature System"** in some of the *most common situations in a routine surveillance scenario*:



Concluding remarks

ReporTree is an **automated and flexible pipeline** that can be used for a **wide variety of species** and that facilitates the **detection of genetic clusters** and their **linkage to epidemiological data**, in a concept aligned with "One Health" perspectives. ReporTree is currently available as a command line tool and can be easily integrated in start-to-end platforms for genomics/epidemiological analysis. For instance, it will be soon integrated in the COHESIVE Information System and in the INSaFLU[6] platform.

ReporTree facilitates and accelerates the production of surveillance-oriented reports, contributing to a sustainable and efficient public health genomics-informed surveillance.

This work was supported by:

- Funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 773830: One Health European Joint Programme (2020-2022)
- National funds through FCT - Foundation for Science and Technology, I.P., in the frame of Individual CEEC 2022.00851.CEECIND/CP1748/CT0001 (2023 onwards)

Bibliography

Zhou et al. (2018) GrapeTree: visualization of core genomic relationships among 100,000 bacterial pathogens. Genome Research, 28(9), 1395–1404.
Balaban et al. (2019) TreeCluster: Clustering biological sequences using phylogenetic trees. PloS One, 14(8), e0221068.
Barker et al. (2018) Rapid identification of stable clusters in bacterial populations using the adjusted Wallace coefficient. In bioRxiv. bioRxiv. https://doi.org/10.1101/299347
Carriço et al. (2006) Illustration of a common framework for relating multiple typing methods by application to macrolide-resistant *Streptococcus pyogenes*. J Clinical Microbiology, 44(7), 2524–2532.
Pinto et al. (2021) *Neisseria gonorrhoeae* clustering to reveal major European whole-genome-sequencing-based genogroups in association with antimicrobial resistance. Microbial Genomics, 7(2).
Borges et al. (2018) INSaFLU: an automated open web-based bioinformatics suite "from-reads" for influenza whole-genome-sequencing-based surveillance. Genome Medicine, 10(1), 46.