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

Verónica Mixão[1], Miguel Pinto[1], João Paulo Gomes[1], Vítor Borges[1]

[1]Bioinformatics Unit, Department of Infectious Diseases National Institute of Health Dr. Ricardo Jorge - Lisbon (Portugal)

SCAN ME

Motivation

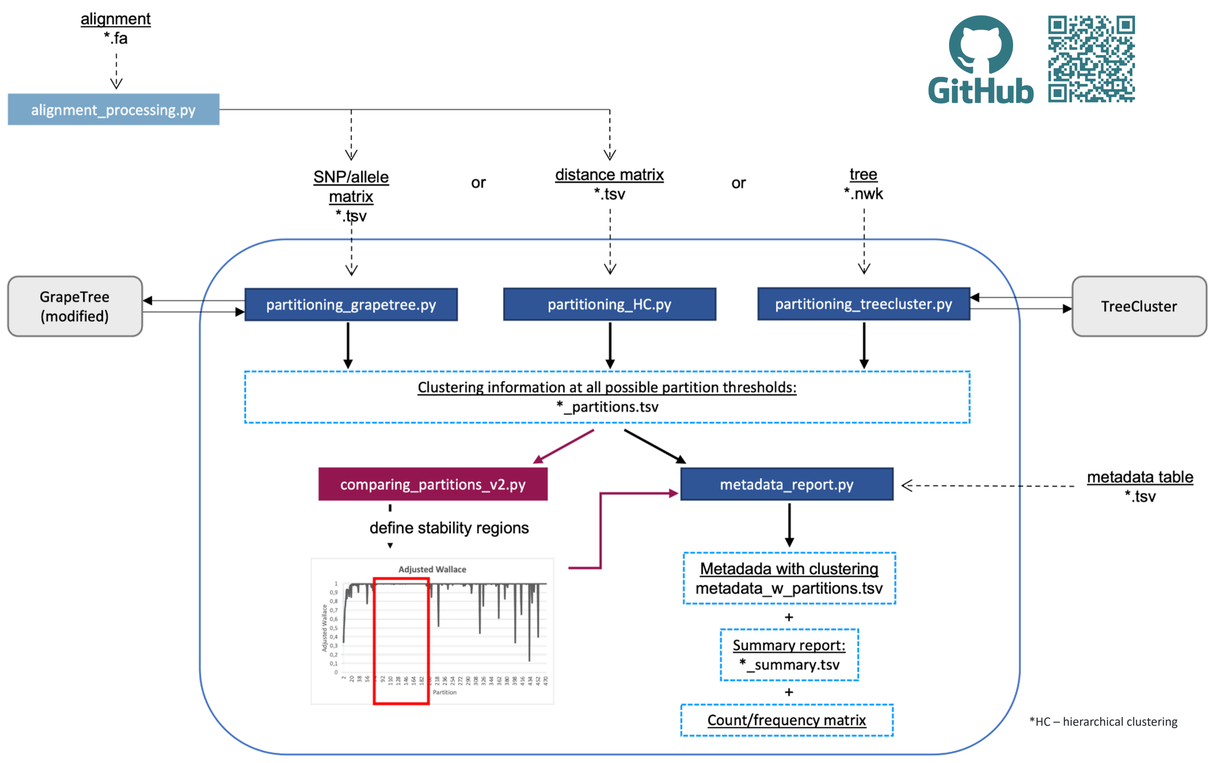
Genomics-informed pathogen surveillance strengthens public health decision-making, thus playing an important role in infectious diseases' prevention and control. A pivotal outcome of genomics surveillance is the identification of pathogen genetic clusters/lineages and their characterization in terms of geotemporal spread or linkage to clinical and demographic data. This task usually relies on the visual exploration of (large) phylogenetic trees (e.g., Minimum Spanning Trees (MST) for bacteria or rooted SNP-based trees for viruses). As this may be a non-trivial, non-reproducible and time-consuming task, **we developed ReporTree, a flexible pipeline that facilitates the detection of genetic clusters and their linkage to epidemiological data.**

ReporTree can help you to…

- obtain **genetic clusters at any threshold level(s)** of a tree, SNP or cg/wgMLST allele matrix, 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.)
- obtain **count/frequency matrices** for the derived genetic clusters or for any other provided grouping variable
- identify **regions of cluster stability** (*i.e.,* threshold ranges in which cluster composition is similar), a key step for nomenclature design

Implementation

ReporTree is an **open-source** tool implemented in **python 3.6** and comprises five main modules orchestrated by *reportree.py* (also available in standalone mode):



ReporTree's modular architecture and flexible functionality makes it compatible and easy to implement with/in several genomic pipelines/platforms

Usage

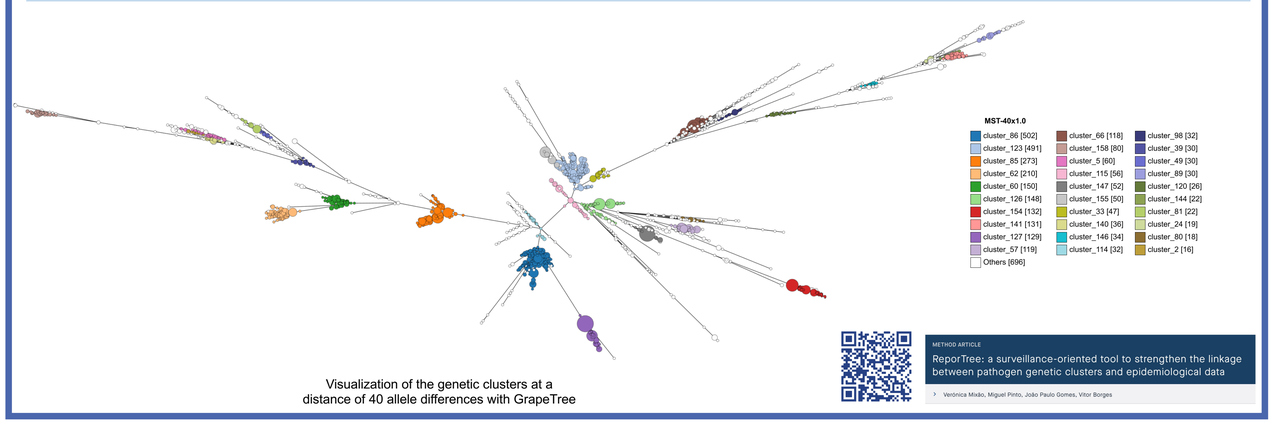
Clustering	Input	Main outputs
Minimum Spanning Tree (using GrapeTree[1])	- **Multiple sequence alignment** (e.g., core SNP alignment) - **SNP/Allele matrix** (e.g., derived from a cg/wgMLST analysis)	• **Genetic clusters** at any (or all) possible distance threshold(s) (partitions table) • **Updated metadata** table with clustering information • **Summary reports with the statistics/ trends for the derived genetic clusters** (e.g., timespan, location, cluster size and composition, age) • **Count/frequency matrices** for the derived genetic clusters or for any other indicated grouping variable • **Regions of cluster stability**[3,4] (a key step in nomenclature design) • **Newick tree** (when applicable)
Hierarchical clustering (using several methods such as single linkage)	- **Multiple sequence alignment** (e.g., core SNP alignment) - **SNP/Allele matrix** (e.g., derived from a cg/wgMLST analysis) - Pairwise **distance matrix**	
Distance between leaves and root or between tree nodes (using TreeCluster[2])	- **Newick tree** (e.g., SNP-scaled tree or dendrogram)	

Validation

We reproduced the extensive genomics analysis of the bacterial pathogen ***Neisseria gonorrhoeae*** performed by Pinto et al., 2021[5]. In this study, 3,791 *N. gonorrhoeae* genomes from isolates collected across Europe were analyzed with a cgMLST approach. Using an **allele matrix** and a **metadata** table as input, with a **single command line** and in 1min 39s, we obtained:

- Regions of cluster stability
- Updated metadata table with clustering information for the first partition of each stability region
- Summary reports for the genetic clusters of the higher and lower levels of stability
- Distribution and occurrence of the genetic determinants of antimicrobial resistance

	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 40 allele differences with GrapeTree

METHOD ARTICLE

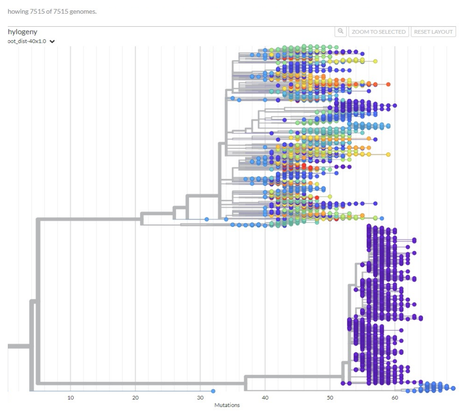
ReporTree: a surveillance-oriented tool to strengthen the linkage between pathogen genetic clusters and epidemiological data

Verónica Mixão, Miguel Pinto, João Paulo Gomes, Vítor Borges

ReporTree and its application to genomics-informed routine surveillance and outbreak investigation

ReporTree was implemented in the routine genomics surveillance of **SARS-CoV-2** in Portugal to speed-up the association between genomic and epidemiological data and the generation of **surveillance-oriented reports**

Inputs: metadata table (tsv) and rooted SNP-scaled tree (newick)



Visualization of the genetic clusters in auspice.us

Main outputs:

- Overview report with the number of sequences, geotemporal distribution and weekly relative frequencies for each lineage/clade at national level
- Lineage/clade weekly relative frequencies at regional level for specific periods of time (e.g., ISO weeks)
- Genetic clusters and their characterization according to any epidemiological/biological relevant indicators included in the metadata

Another direct application is the analysis of cg/wgMLST data for **outbreak investigation**, namely, for foodborne bacterial pathogens. In *ReporTree's wiki* we provide a simulated example in which ReporTree automatically extracts and reports genetic clusters of ***Listeria monocytogenes*** at high-resolution levels commonly used for outbreak detection.

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 easily be integrated in surveillance-oriented workflows for genomics / epidemiological data analysis (for instance, it will be integrated in INSaFLU[6]), thus contributing to a sustainable and efficient public health genomics-informed pathogen surveillance.

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

Funding:

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.

https://onehealthejp.eu/jrp-beone/

References

[1] Zhou et al. (2018) GrapeTree: visualization of core genomic relationships among 100,000 bacterial pathogens. Genome Research, 28(9), 1395–1404.
[2] Balaban et al. (2019) TreeCluster: Clustering biological sequences using phylogenetic trees. PloS One, 14(8), e0221068.
[3] 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
[4] 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.
[5] 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).
[6] 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.