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

Verónica Mixão
Bioinformatics Unit, Department of Infectious Diseases, National Institute of Health Dr. Ricardo Jorge, Av. Padre Cruz, 1600-609 Lisbon, Portugal https://orcid.org/0000-0001-6669-0161

Miguel Pinto
Bioinformatics Unit, Department of Infectious Diseases, National Institute of Health Dr. Ricardo Jorge, Av. Padre Cruz, 1600-609 Lisbon, Portugal

João Paulo Gomes
Bioinformatics Unit, Department of Infectious Diseases, National Institute of Health Dr. Ricardo Jorge, Av. Padre Cruz, 1600-609 Lisbon, Portugal

Vitor Borges (vitor.borges@insa.min-saude.pt)
Bioinformatics Unit, Department of Infectious Diseases, National Institute of Health Dr. Ricardo Jorge, Av. Padre Cruz, 1600-609 Lisbon, Portugal

Method Article

Keywords: genetic clusters, genomics-informed surveillance, surveillance-oriented reports

Posted Date: February 28th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1404655/v1

Abstract

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 and their characterization in terms of geotemporal spread or linkage to clinical and demographic data. This task often consists of the visual exploration of (large) phylogenetic trees and associated metadata, being time consuming and difficult to reproduce. For this reason, we developed ReporTree, an automated surveillance-oriented pipeline that allows diving into the complexity of pathogen diversity to rapidly identify genetic clusters at any (or all) distance threshold(s) of a tree, identify regions of cluster stability (key step in nomenclature design), and further characterize them according to any relevant feature, such as timespan, geography or clinical status. To validate ReporTree, we reproduced a large-scale study on genetic clustering and linkage to antibiotic resistance data in *Neisseria gonorrhoeae*. Also, we show how this tool can be a useful asset in genomics-informed routine surveillance and outbreak detection of a wide variety of species, providing examples for SARS-CoV-2 virus and the foodborne bacterial pathogen *Listeria monocytogenes*. In summary, ReporTree facilitates and accelerates the production of surveillance-oriented reports, thus contributing to a sustainable and efficient public health genomics-informed pathogen surveillance.

Availability and implementation: ReporTree is implemented in python 3.6 and is freely available at https://github.com/insapathogenomics/ReporTree.

Motivation

The implementation of genomics-informed surveillance systems able to track the circulation of pathogens and monitor their clinical and epidemiologically relevant features is essential for infectious diseases' prevention and control and for a more informed public health decision-making. Whole-genome sequencing (WGS) is the method with the highest resolution to discriminate and classify microorganisms (either at inter- or intra-species level) based on their genetic relatedness, thus playing a pivotal role in pathogen surveillance and research activities.

Several bioinformatics solutions for the analysis of WGS data are currently available, with most workflows for genetic clustering determination ending-up in the same key output: a phylogenetic tree. It usually corresponds to a Minimum Spanning Tree (MST) reflecting the allele distances that result from a core genome- (cg) or whole genome- (wg) Multiple Locus Sequence Type (MLST) analysis (commonly used approach for bacterial pathogens (Jolley & Maiden, 2014)), or to a rooted tree reflecting the Single Nucleotide Polymorphism (SNP) distances that result from a multiple sequence alignment (e.g., as routinely applied for viruses (Wohl et al., 2016), such as SARS-CoV-2). Subsequently, the identification and characterization of epidemiologically/biologically relevant genetic clusters (e.g., clusters of outbreak-related strains) is often performed through visual inspection of these trees in association with the samples' metadata, taking advantage of robust visualization features, such as those provided by PHYLOViZ (Ribeiro-Gonçalves et al., 2016), GrapeTree (Zhou et al., 2018), Nextstrain (Hadfield et al.,

2018) or Microreact (Argimón et al., 2016). In this context, there is a continuous scientific effort to automate the identification of clusters at specific genetic thresholds (Balaban et al., 2019; Dallman et al., 2018; Deneke et al., 2021; Ragonnet-Cronin et al., 2013) and develop dynamic cluster/lineage nomenclature systems, such as the Pango system for SARS-CoV-2 (Rambaut et al., 2020), or the bacteria-oriented “SNP address” of SnapperDB workflow (Dallman et al., 2018) and the “allele hash” of chewieSnake workflow (Deneke et al., 2021). Still, the field would benefit from the development of automated and more flexible tools that can be used for a wide variety of species, not only to facilitate the detection of genetic clusters at any (or all) distance thresholds of a tree, but also to automatically characterize them based on the available metadata variables of interest.

Here, we present ReporTree, an automated surveillance-oriented resource that allows diving into the complexity of pathogen diversity to rapidly identify genetic clusters at any threshold(s) of a tree (both MST and rooted tree) and further characterize them according to any relevant epidemiological indicator in a reproducible manner.

Implementation

ReporTree is implemented in python 3.6 and comprises four main modules orchestrated by *reportree.py* (Fig. 1) and also available in standalone mode:

a) *partitioning_grapetree.py* - this script takes as input an allele matrix and uses a modified version of GrapeTree (Zhou et al., 2018) (available at https://github.com/insapathogenomics/GrapeTree) to obtain a MST. Then, it identifies genetic clusters at any user-defined partition thresholds or, alternatively, at all possible thresholds of the tree. Additionally, this script can filter the allele matrix based on the number of loci called in a cg/wgMLST analysis and/or on the metadata information according to the filter parameters specified by the user, thus generating the MST for a sample subset. The output is a “partitions table” with clustering information for each sample at all the distance thresholds analyzed.

b) *partitioning_treecluster.py* - this script takes as input a newick (rooted) tree and takes advantage of TreeCluster (Balaban et al., 2019) to determine genetic clusters at user-defined combination(s) of any clustering method(s) and any (or all) distance threshold(s) of the tree. The output is a “partitions table” with clustering information for each sample at all the distance thresholds analyzed.
c) *comparing_partitions_v2.py* - this script represents a new version of the code that is the basis of the Comparing Partitions tool (http://www.comparingpartitions.info) (Carriço et al., 2006). This new version takes as input a partitions table with clustering information at all possible thresholds of a tree. Then, it compares the clustering information at different partitions using the neighborhood Adjusted Wallace coefficient (nAWC) (Carriço et al., 2006; Severiano et al., 2011), and determines regions of cluster stability (i.e. subsequent partition thresholds in which clustering information is similar) based on a previously described approach (Barker et al., 2018; Llarena et al., 2018). If the user requests, ReporTree uses this information to decide on the thresholds for which cluster characterization must be performed.

d) *metadata_report.py* - this script takes as input a metadata table and, according to the user's specifications, outputs summary reports for the requested metadata columns (e.g., lineage) with summary/trends of other metadata features (e.g., vaccination status, timespan, antibiotic resistance phenotype, etc.). Furthermore, it can output relative frequency or count matrices for a given metadata variable (e.g., frequency of each lineage per week). If besides the metadata table the user also provides a "partitions table", this script can perform cluster characterization according to any relevant epidemiological indicator present in the metadata.

In summary, ReporTree is a flexible tool able to identify clustering information at any partition thresholds either for species that require a cg/wgMLST analysis or for those that rely on (rooted) tree reconstruction (Fig. 1).

Validation

Reproducing a large-scale study on genetic clustering and linkage to antibiotic resistance data in Neisseria gonorrhoeae

Our team has recently performed an extensive genomics analysis of the bacterial pathogen *Neisseria gonorrhoeae* (Pinto et al., 2021). In this study, 3,791 *N. gonorrhoeae* genomes from isolates collected across Europe were analyzed with a cgMLST approach. Genetic clusters were determined with the goeBURST algorithm implemented in PHYLOViZ (Francisco et al., 2009, 2012; Nascimento et al., 2017; Ribeiro-Gonçalves et al., 2016) for all possible allelic distance thresholds (partitions). Cluster concordance between subsequent distance thresholds was assessed with the nAWC in order to determine regions of cluster stability (Barker et al., 2018; Carriço et al., 2006; Llarena et al., 2018; Severiano et al., 2011) that were used for nomenclature purposes and identification of genogroups. The association between metadata and genetic clusters was then performed by time-consuming table handling with a spreadsheet program. This corresponded to a non-automated workflow and, in the particular case of the cluster congruence analysis and the integration of genetic and clinically or epidemiologically relevant data, it represented a highly demanding process difficult to be applied in real-time pathogen surveillance. As such, to validate ReporTree and demonstrate how it can enhance bacterial pathogens' surveillance and research, we used the same dataset as in the previous study (Pinto et al., 2021) and attempted to reproduce the main study outputs with this tool. As shown at https://github.com/insapathogenomics/ReporTree/wiki/, using as input the allele matrix with 822 loci (available at https://zenodo.org/record/3946223#.YhTKQy8qKrw) and the associated metadata (available in Supplementary material 1 of (Pinto et al., 2021)), ReporTree automatically identified the genetic clusters at all possible partition thresholds of the generated MST and identified the same regions of cluster stability as Pinto et al.. Moreover, it provided an updated metadata table with clustering information at the first partition of each stability region, which could be used as input for visualization in GrapeTree (Zhou et al., 2018). Furthermore, summary reports with statistics/trends associated with each genetic cluster of low and high levels of stability (i.e. 40 allele differences at the lower level and 79 allele differences at the higher level, similarly to what was found by Pinto et al.) were reported. Finally,

ReporTree was able to associate and report the distribution of genetic determinants of antimicrobial resistance in *N. gonorrhoeae* for the different genetic clusters. Importantly, this example allowed a clear validation of the tool by rigorously reproducing the data presented, for example, in Figs. 1a, 1b and 3 and in Tables 1 and 2 of (Pinto et al., 2021). All these outputs (and additional ones) are available for consultation at https://github.com/insapathogenomics/ReporTree/tree/main/examples/Neisseria. Noteworthy, this proof of concept was made with a single command line that ran for approximately 1min 39sec in a laptop [Intel Core i5(R)] with 16 GB of RAM.

ReporTree and its application to genomics-informed routine surveillance (e.g., SARS-CoV-2) and outbreak detection (e.g., Listeria monocytogenes**)**

Genomics-informed surveillance of SARS-CoV-2 has had an important role in worldwide public health and political decision-making in the last two years. In Portugal, weekly reports of nationwide sequencing surveys are provided to public health authorities and general public describing important indicators and trends of the evolution and geotemporal spread of the virus (https://insaflu.insa.pt/covid19/). Therefore, after ReporTree validation, we implemented this tool in the routine genomics surveillance of SARS-CoV-2 in the country with the objective of speeding-up the association between genomic and epidemiological data and the generation of the surveillance-oriented reports. For instance, besides its comprehensive usage for monitoring the relative frequency of variants of concern (VOCs) at regional and national levels, ReporTree is applied to identify clusters of high-closely related viruses (e.g., using TreeCluster (Balaban et al., 2019) max-clade or avg-clade models at high resolution levels) that may represent local transmission networks or even super-spreading events. An example of ReporTree application in the context of SARS-CoV-2 genomic epidemiology is provided in https://github.com/insapathogenomics/ReporTree/wiki/.

ReporTree can be useful to a broad spectrum of species. One of the most direct and intuitive applications is the analysis of cg/wgMLST data for outbreak investigation, namely for foodborne bacterial pathogens, as this subtyping method delivers sufficiently high resolution and epidemiological concordance (Nadon et al., 2017). In ReporTree wiki (https://github.com/insapathogenomics/ReporTree/wiki/), it is provided a simple simulated example in which, with a single command line, ReporTree builds a MST from cgMLST data and automatically extracts and reports genetic clusters of *Listeria monocytogenes* at high resolution levels commonly used for outbreak detection (< 5 and < 8 allelic differences, (Van Walle et al., 2018)), as routinely performed in Portugal.

In both examples, ReporTree is a useful asset by rapidly generating summary statistics/trends for key surveillance metrics, such as the relative frequency of the different (sub-)lineages/clusters circulating in the country over time. Noteworthy, ReporTree was designed to provide a high degree of flexibility, allowing, for example, the rapid production of summary and count/frequency reports. While summary reports include the distribution of any (and as many) user-specified variable of interest (e.g. vaccination status, source, country, timespan, etc.) over any user-specified grouping variable (e.g. lineage/cluster or age), count/relative frequency reports include the distribution of one grouping variable over any other user-specified variable (e.g. lineage/cluster per week, or clade per country). When the time variable "date"

is provided in the metadata, ReporTree automatically infers other time units (ISO week and ISO year) and metrics (e.g., cluster timespan) relevant for surveillance purposes. Furthermore, ReporTree allows the application of filters in the metadata table to select the samples that will be included in the report without the need of generating a new metadata table.

Concluding Remarks

ReporTree represents 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. Here, we presented the proof of concept of this tool, showing its ability to report a comprehensive WGS-based genogroup assignment for *N. gonorrhoeae*, based on the identification of the discriminatory genetic thresholds reflecting cluster stability, and the rapid correlation of these genogroups (representing main circulating lineages) with any data of interest, such as antimicrobial resistance data. Furthermore, we have shown how its flexibility contributed to speed up SARS-CoV-2 and *L. monocytogenes* genomics-informed surveillance in Portugal, facilitating and accelerating the production of surveillance-oriented reports. Although ReporTree is currently available as a command line tool, this resource can easily be integrated in surveillance-oriented platforms for genomics / epidemiological analysis (for instance, it will be soon integrated in INSaFLU (Borges et al., 2018)), thus contributing to a sustainable and efficient public health genomics-informed pathogen surveillance.

Declarations

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.

Conflict of interest

None of the authors has a conflict of interest to declare.

References

1. Argimón, S., Abudahab, K., Goater, R. J. E., Fedosejev, A., Bhai, J., Glasner, C., Feil, E. J., Holden, M. T. G., Yeats, C. A., Grundmann, H., Spratt, B. G., & Aanensen, D. M. (2016). Microreact: visualizing and sharing data for genomic epidemiology and phylogeography. Microbial Genomics, *2*(11), e000093.
2. Balaban, M., Moshiri, N., Mai, U., Jia, X., & Mirarab, S. (2019). TreeCluster: Clustering biological sequences using phylogenetic trees. PloS One, *14*(8), e0221068.
3. Barker, D. O. R., Carriço, J. A., Kruczkiewicz, P., Palma, F., Rossi, M., & Taboada, E. N. (2018). Rapid identification of stable clusters in bacterial populations using the adjusted Wallace coefficient. In

bioRxiv. bioRxiv. https://doi.org/10.1101/299347

4. Borges, V., Pinheiro, M., Pechirra, P., Guiomar, R., & Gomes, J. P. (2018). INSaFLU: an automated open web-based bioinformatics suite “from-reads” for influenza whole-genome-sequencing-based surveillance. Genome Medicine, *10*(1), 46.
5. Carriço, J. A., Silva-Costa, C., Melo-Cristino, J., Pinto, F. R., de Lencastre, H., Almeida, J. S., & Ramirez, M. (2006). Illustration of a common framework for relating multiple typing methods by application to macrolide-resistant Streptococcus pyogenes. Journal of Clinical Microbiology, *44*(7), 2524–2532.
6. Dallman, T., Ashton, P., Schafer, U., Jironkin, A., Painset, A., Shaaban, S., Hartman, H., Myers, R., Underwood, A., Jenkins, C., & Grant, K. (2018). SnapperDB: a database solution for routine sequencing analysis of bacterial isolates. Bioinformatics, *34*(17), 3028–3029.
7. Deneke, C., Uelze, L., Brendebach, H., Tausch, S. H., & Malorny, B. (2021). Decentralized Investigation of Bacterial Outbreaks Based on Hashed cgMLST. Frontiers in Microbiology, *12*, 649517.
8. Francisco, A. P., Bugalho, M., Ramirez, M., & Carriço, J. A. (2009). Global optimal eBURST analysis of multilocus typing data using a graphic matroid approach. BMC Bioinformatics, *10*, 152.
9. Francisco, A. P., Vaz, C., Monteiro, P. T., Melo-Cristino, J., Ramirez, M., & Carriço, J. A. (2012). PHYLOViZ: phylogenetic inference and data visualization for sequence based typing methods. BMC Bioinformatics, *13*, 87.
10. Hadfield, J., Megill, C., Bell, S. M., Huddleston, J., Potter, B., Callender, C., Sagulenko, P., Bedford, T., & Neher, R. A. (2018). Nextstrain: real-time tracking of pathogen evolution. Bioinformatics, *34*(23), 4121–4123.
11. Jolley, K. A., & Maiden, M. C. J. (2014). Using multilocus sequence typing to study bacterial variation: prospects in the genomic era. Future Microbiology, *9*(5), 623–630.
12. Llarena, A.-K., Ribeiro-Gonçalves, B. F., Nuno Silva, D., Halkilahti, J., Machado, M. P., Da Silva, M. S., Jaakkonen, A., Isidro, J., Hämäläinen, C., Joenperä, J., Borges, V., Viera, L., Gomes, J. P., Correia, C., Lunden, J., Laukkanen-Ninios, R., Fredriksson-Ahomaa, M., Bikandi, J., Millan, R. S., … Rossi, M. (2018). INNUENDO: A cross-sectoral platform for the integration of genomics in the surveillance of food-borne pathogens. EFSA Supporting Publications, *15*(11). https://doi.org/10.2903/sp.efsa.2018.en-1498
13. Nadon, C., Van Walle, I., Gerner-Smidt, P., Campos, J., Chinen, I., Concepcion-Acevedo, J., Gilpin, B., Smith, A. M., Man Kam, K., Perez, E., Trees, E., Kubota, K., Takkinen, J., Nielsen, E. M., Carleton, H., & FWD-NEXT Expert Panel. (2017). PulseNet International: Vision for the implementation of whole genome sequencing (WGS) for global food-borne disease surveillance. Euro Surveillance: Bulletin Europeen Sur Les Maladies Transmissibles = European Communicable Disease Bulletin, *22*(23). https://doi.org/10.2807/1560-7917.ES.2017.22.23.30544
14. Nascimento, M., Sousa, A., Ramirez, M., Francisco, A. P., Carriço, J. A., & Vaz, C. (2017). PHYLOViZ 2.0: providing scalable data integration and visualization for multiple phylogenetic inference methods. Bioinformatics, *33*(1), 128–129.

15. Pinto, M., Borges, V., Isidro, J., Rodrigues, J. C., Vieira, L., Borrego, M. J., & Gomes, J. P. (2021). clustering to reveal major European whole-genome-sequencing-based genogroups in association with antimicrobial resistance. Microbial Genomics, *7*(2). https://doi.org/10.1099/mgen.0.000481
16. Ragonnet-Cronin, M., Hodcroft, E., Hué, S., Fearnhill, E., Delpech, V., Brown, A. J. L., Lycett, S., & UK HIV Drug Resistance Database. (2013). Automated analysis of phylogenetic clusters. BMC Bioinformatics, *14*, 317.
17. Rambaut, A., Holmes, E. C., O'Toole, Á., Hill, V., McCrone, J. T., Ruis, C., du Plessis, L., & Pybus, O. G. (2020). A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. Nature Microbiology, *5*(11), 1403–1407.
18. Ribeiro-Gonçalves, B., Francisco, A. P., Vaz, C., Ramirez, M., & Carriço, J. A. (2016). PHYLOViZ Online: web-based tool for visualization, phylogenetic inference, analysis and sharing of minimum spanning trees. Nucleic Acids Research, *44*(W1), W246–W251.
19. Severiano, A., Pinto, F. R., Ramirez, M., & Carriço, J. A. (2011). Adjusted Wallace coefficient as a measure of congruence between typing methods. Journal of Clinical Microbiology, *49*(11), 3997–4000.
20. Van Walle, I., Björkman, J. T., Cormican, M., Dallman, T., Mossong, J., Moura, A., Pietzka, A., Ruppitsch, W., Takkinen, J., & European Listeria WGS typing group. (2018). Retrospective validation of whole genome sequencing-enhanced surveillance of listeriosis in Europe, 2010 to 2015. Euro Surveillance: Bulletin Europeen Sur Les Maladies Transmissibles = European Communicable Disease Bulletin, *23*(33). https://doi.org/10.2807/1560-7917.ES.2018.23.33.1700798
21. Wohl, S., Schaffner, S. F., & Sabeti, P. C. (2016). Genomic Analysis of Viral Outbreaks. Annual Review of Virology, *3*(1), 173–195.
22. Zhou, Z., Alikhan, N.-F., Sergeant, M. J., Luhmann, N., Vaz, C., Francisco, A. P., Carriço, J. A., & Achtman, M. (2018). GrapeTree: visualization of core genomic relationships among 100,000 bacterial pathogens. Genome Research, *28*(9), 1395–1404.

Figures

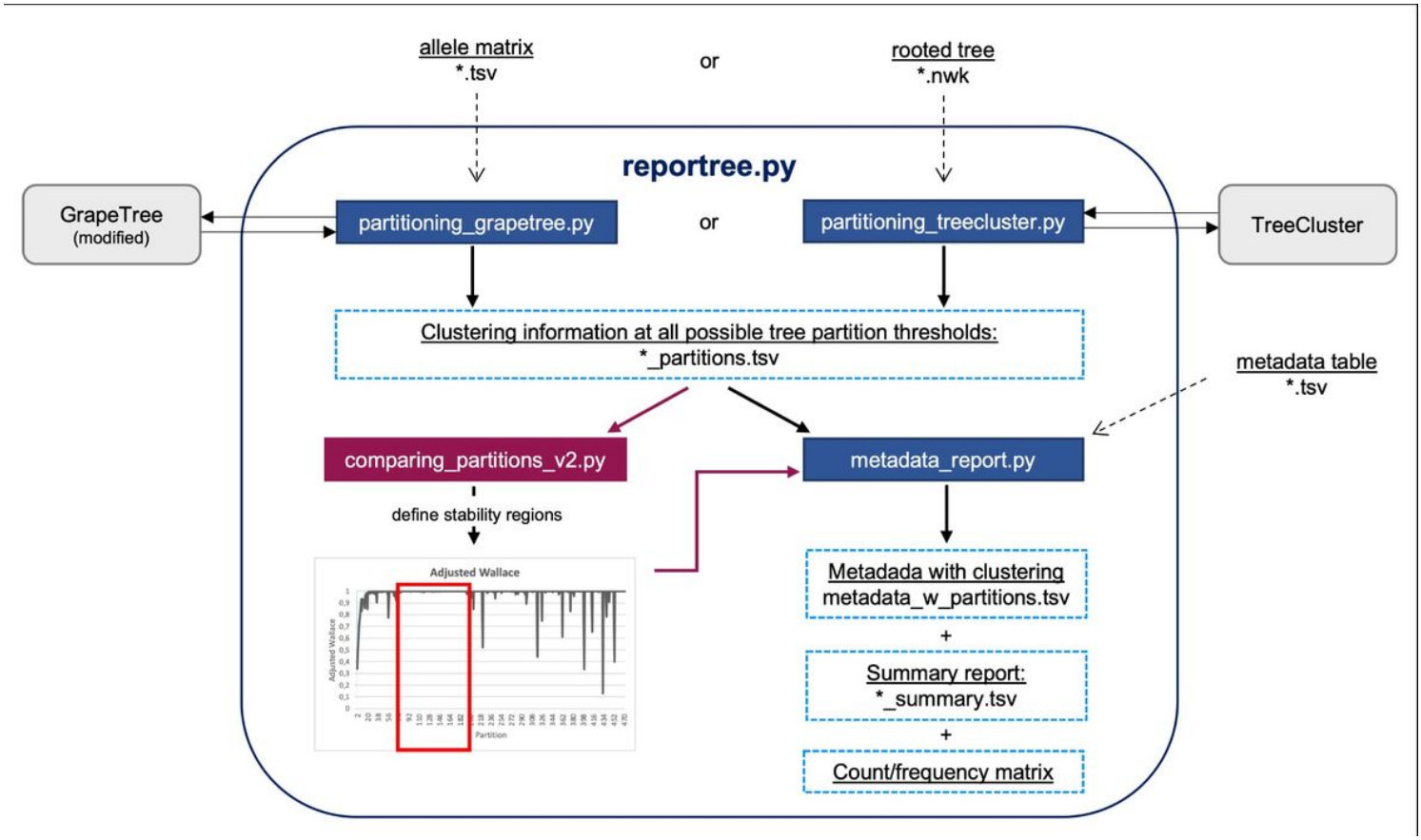

Figure 1

Schematic representation of ReporTree pipeline. ReporTree comprises four modules, which are highlighted in dark blue (core modules) and dark pink (extra module that can only be run if the partitions table includes clustering information at all possible thresholds). Dashed boxes indicate the outputs of each module. More details and examples of usage can be found in ReporTree github repository (https://github.com/insapathogenomics/ReporTree) and wiki (https://github.com/insapathogenomics/ReporTree/wiki).