




Visual linkage and interactive features of Evidente for an enhanced analysis of SNP-based phylogenies

M. Witte Paz¹ , T. Harbig¹ , D. Varga¹, E. Kränzle¹, and K. Nieselt¹ 

¹ University of Tübingen, Institute for Bioinformatics and Medical Informatics, Tübingen, Germany

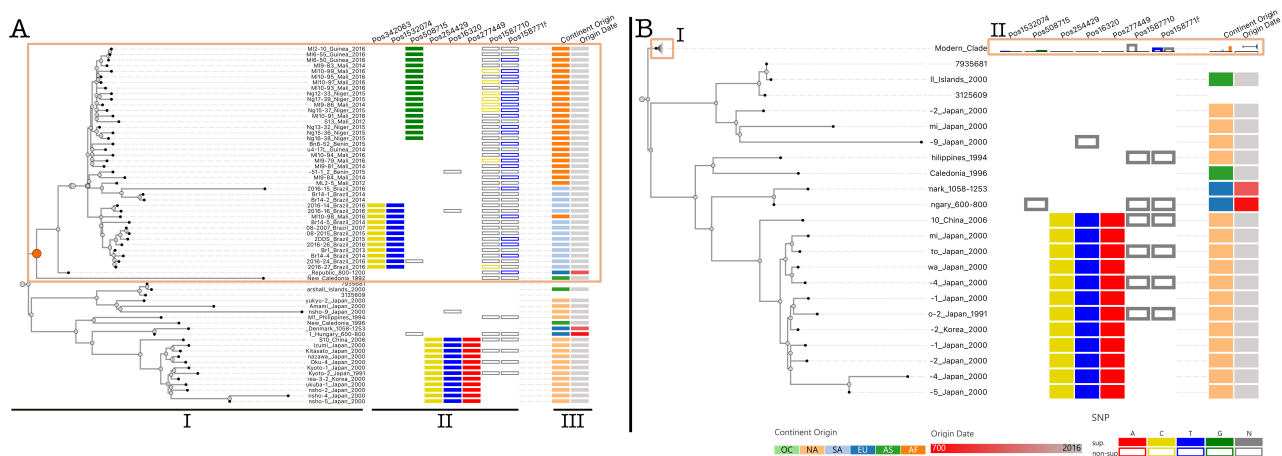


Figure 1: Genomic data of *Mycobacterium leprae* [SAKK* 18] visualized with *Evidente*. Orange rectangles indicate different views of the same data. (A) Full view of the phylogenetic tree (A.I), using a horizontal dendrogram, aligned to two heat maps. The orange node indicates the root of the collapsed clade in (B). The first heat map (A.II) shows multiple SNPs across the strains. Full rectangles encode unique SNPs for a clade (supporting SNPs), while SNPs found in many clades are visualized as colored frames (non-supporting SNPs). The metadata heatmap (A.III) shows two columns, one for continent of origin (categorical data) and one for date of origin (numerical data). (B) Overview of the tree with a collapsed clade indicated via a glyph (B.I). SNP data and metadata are aggregated (B.II) using bar charts (SNPs and categorical metadata) and boxplots (numerical metadata).

Abstract

Phylogenetic trees of a set of bacterial strains are often used to analyze their evolutionary relationships and they are commonly based on genomic features, such as single nucleotide polymorphisms (SNPs). *Evidente* - a recently published tool - provides visual and analytical linkage across a phylogenetic tree, SNP data and metadata, and integrates them into one interactive visual analytics platform. In contrast to other approaches, *Evidente* shows how SNPs agree with the tree structure. *Evidente* is part of the TueVis server (<https://evidente-tuevis.cs.uni-tuebingen.de/>). Here, we give an overview of the tasks supported by *Evidente*. The version of *Evidente* described in the publication can seamlessly visualize up to 150 strains. Thus, we introduce further enhancements for larger trees, such as data-driven aggregation and semantic zooming.

1. Introduction

A common question in biology is how individual strains of a bacterial species are related, since this can help to understand, for example, how specific strains developed their virulence or acquired their antibiotic resistance. One approach to answer this question is to compare the genomes of the strains and to identify single-point mu-

tations also called single-nucleotide polymorphisms (SNPs). These variant sites across strains can then be used to reconstruct their evolutionary history in the form of a phylogenetic tree. However, phylogenetic reconstruction methods typically return only the structure of the tree without providing a linkage between the underlying genomic data and the tree. This linkage would allow the identification

© 2023 The Authors.

Proceedings published by Eurographics - The European Association for Computer Graphics.

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

of patterns across the data sources. Therefore, improving the explainability between these data types was one of the main goals in developing a visual analytics tool (Task T_1). Moreover, users rarely analyze the evolutionary history independently of other metadata, for example, the antibiotic resistance level or the geographic origin of the strains. Hence, it might be important to identify possible patterns of SNPs or metadata specific for sub-clusters of the tree, also known as clades (Task T_2), or to find non-clade specific patterns correlating with metadata across the phylogenetic tree (Task T_3). Lastly, these analyses often need to compare many strains simultaneously. For this, filtering and summarization techniques should be included to allow the interaction with such datasets (Task T_4).

Though different tools already visualize phylogenetic trees together with further data [SGL*19, KBC*17, HCG*18], they cannot be directly used to identify the agreement of the phylogenetic tree with the genomic data. To cover this gap in the visualization, the visual analytics tool *Evidente* (Efficient VISual analytics tool for Data ENrichment in phylogenetic TreEs) was recently published in *Bioinformatic Advances* [WPHN22]. By computing the specificity of a SNP with respect to the clades, *Evidente* is able to improve the explainability of the evolutionary history. *Evidente* classifies the SNPs into two categories: SNPs that are unique for a clade (supporting SNPs) and those distributed across many clades (non-supporting SNPs) (see [WPHN22] for details). Furthermore, SNPs that could not be accurately identified (unresolved bases) are also classified as non-supporting SNPs. However, the key feature of *Evidente* is the integration of the SNP specificity in a single interactive visualization that simultaneously shows the phylogenetic tree and the SNP data, as well as metadata. With interactive filtering and aggregation features, the analysis of datasets with a few hundred samples has been made possible.

2. Visual Analytics Interface

To allow the exploration of data of diverse nature, *Evidente* aligns a tree view with tabular data in a single-page interface (Fig. 1), following a similar approach as described in *Jupiter* [NSL18]. In *Evidente*, the tree is a horizontal dendrogram visualizing the phylogenetic tree (Fig. 1A.I) via the library *PhyloTree* [SWP18]. The tabular data is visualized via two heat maps, one for SNP data and one for metadata. In the heat map of the SNP data (Fig. 1A.II), each column represents one position in the reference genome, a rectangle encodes the presence of a SNP (i.e. the absence of a rectangle refers to the sample containing the reference allele), while its color encodes for the SNP's allele. Supporting SNPs are visualized via full rectangles, while colored frames with white fills represent non-supporting SNPs. The user can access and visualize the SNPs that are unique for each clade via the root of each subtree (Task T_1 and T_2).

In the metadata heat map each column represents one instance of the metadata (Fig. 1A.III), and the color scales vary depending on their type. Numerical data are visualized using a linear continuous color scale, while ordinal data is encoded by a sequential color scale. For categorical data, a set of 20 colors is defined by default [BOH11]. Aligning the tree, SNP data, and metadata allows the user to identify patterns, either within specific clades (Task T_2) or across the phylogenetic tree (Task T_3).

Evidente also provides interactive features to enhance the exploration of large datasets (Task T_4). The leaves of the tree can

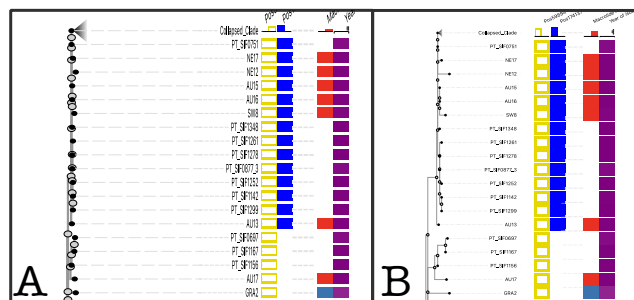


Figure 2: Zoom in on a region of the visualization of genomic data of *Treponema pallidum* [PDSBG*22] visualized with *Evidente*. For the full phylogenetic tree see Sup. Figure S1. (A) Geometric zooming implemented in the first version. (B) New implementation of semantic zooming

be filtered out via the dendrogram or via metadata-based filters. This reduces the number of visualized elements, but it also excludes nodes from the analysis. Hence, users can choose to collapse specific clades (Fig. 1B). These are indicated by a glyph (Fig. 1B.I) and the corresponding data is summarized (Fig. 1B.II).

Evidente is available at the TueVis server (<https://evidente-tuevis.cs.uni-tuebingen.de/>). With its ability to integrate different data types into a holistic view, the version of *Evidente* introduced in the original publication has already facilitated the exploration of phylogenies together with SNP data and metadata.

3. Enhanced Features to Interact with Large Trees

For a complete view of the evolutionary history, the initially published version of *Evidente* showed the entire phylogenetic tree. Therefore, this approach is most effective for trees consisting of less than 150 samples. For larger trees, the height of rows in the heat map can become too small for effective interaction with the data. Since we aim to increase the scope of *Evidente* to analyze even larger data sets, we have implemented methods to automatically collapse clades depending on their depth, their SNP content, or via metadata (see Suppl. Video 1). This enhances the visualization, since collapsed clades display SNPs and metadata via summary visualizations. To further facilitate the analysis of large trees, we have replaced the geometric zooming of *Evidente* with a semantic zooming approach. While the geometric zoom increased the size of the pixels, the semantic zooming increases the distances within the elements for an enhanced exploration (see Fig. 2 and Suppl. Video 2). In a future release of *Evidente* we plan to implement further enhancements, such as aggregation of heat map columns of identical SNP patterns. Moreover, a visual distinction of more specific SNP classifications (monophyletic, paraphyletic, and polyphyletic) will be implemented.

Acknowledgements

This project has been supported by infrastructural funding from the Cluster of Excellence EXC 2124 ‘Controlling Microbes to Fight Infections’ [project ID 390838134 to MWP and KN] and from TRR261 [project ID 398967434 to TH], both from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation).

References

- [BOH11] BOSTOCK M., OGIEVETSKY V., HEER J.: D³ data-driven documents. *IEEE Transactions on Visualization and Computer Graphics* 17, 12 (2011), 2301–2309. [2](#)
- [HCG*18] HADFIELD J., CROUCHER N. J., GOATER R. J., ABUDAHAB K., AANENSEN D. M., HARRIS S. R.: Phandango: an interactive viewer for bacterial population genomics. *Bioinformatics* 34, 2 (2018), 292–293. [2](#)
- [KBC*17] KREFT Ł., BOTZKI A., COPPENS F., VANDEPOELE K., VAN BEL M.: PhyD3: a phylogenetic tree viewer with extended phyloXML support for functional genomics data visualization. *Bioinformatics* 33, 18 (2017), 2946–2947. [2](#)
- [NSL18] NOBRE C., STREIT M., LEX A.: Juniper: A tree+ table approach to multivariate graph visualization. *IEEE transactions on visualization and computer graphics* 25, 1 (2018), 544–554. [2](#)
- [PDSBG*22] PLA-DÍAZ M., SÁNCHEZ-BUSÓ L., GIACANI L., ŠMAJS D., BOSSHARD P. P., BAGHERI H. C., SCHUENEMANN V. J., NIESELT K., ARORA N., GONZÁLEZ-CANDELAS F.: Evolutionary processes in the emergence and recent spread of the syphilis agent, *Treponema pallidum*. *Molecular biology and evolution* 39, 1 (2022), msab318. [2](#)
- [SAKK*18] SCHUENEMANN V. J., AVANZI C., KRAUSE-KYORA B., SEITZ A., HERBIG A., INSKIP S., BONAZZI M., REITER E., URBAN C., DANGVARD PEDERSEN D., ET AL.: Ancient genomes reveal a high diversity of *Mycobacterium leprae* in medieval Europe. *PLOS Pathogens* 14, 5 (2018), e1006997. [1](#)
- [SGL*19] SUBRAMANIAN B., GAO S., LERCHER M. J., HU S., CHEN W.-H.: Evolvview v3: a webserver for visualization, annotation, and management of phylogenetic trees. *Nucleic Acids Research* 47, W1 (2019), W270–W275. [2](#)
- [SWP18] SHANK S. D., WEAVER S., POND S. L. K.: phylotree.js - a JavaScript library for application development and interactive data visualization in phylogenetics. *BMC Bioinformatics* 19, 1 (2018), 276. [2](#)
- [WPHN22] WITTE PAZ M., HARBIG T. A., NIESELT K.: Evidente—a visual analytics tool for data enrichment in SNP-based phylogenetic trees. *Bioinformatics Advances* 2, 1 (10 2022). vbac075. URL: <https://doi.org/10.1093/bioadv/vbac075>, doi: [10.1093/bioadv/vbac075](https://doi.org/10.1093/bioadv/vbac075). [2](#)