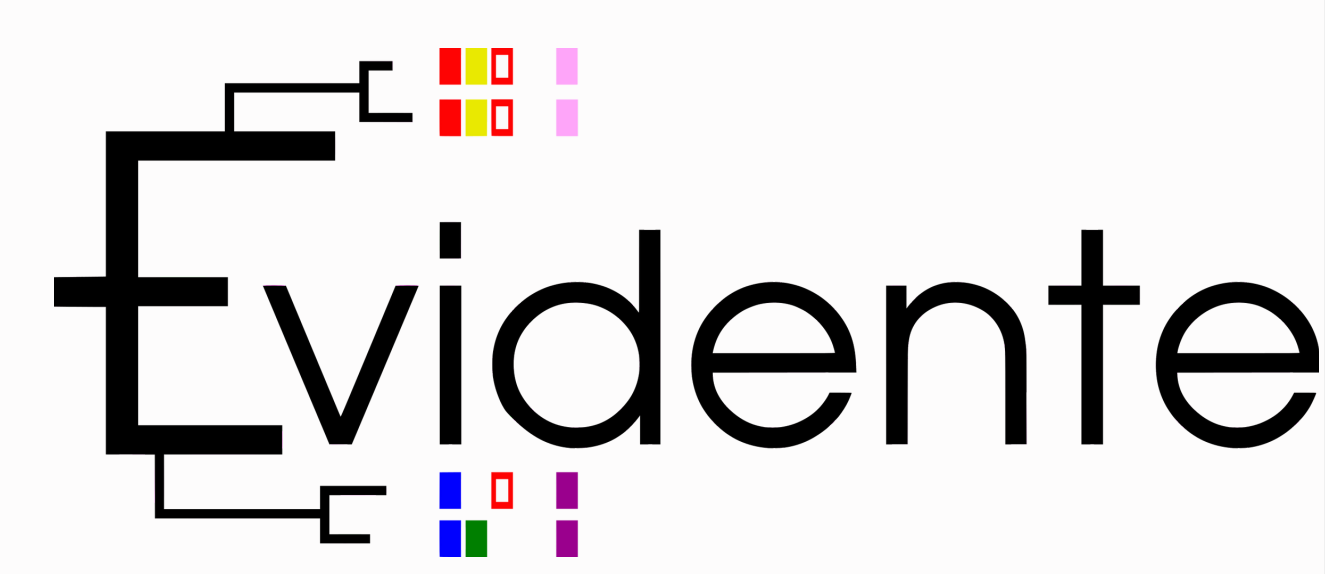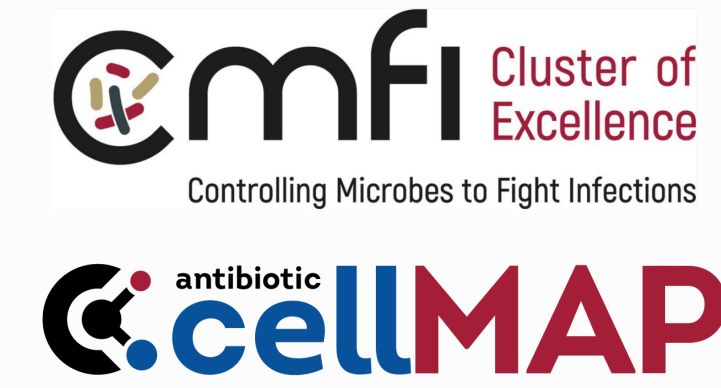# Enhanced analysis of large SNP-based phylogenies via visual linkage and interactive features within Evidente

M. Witte Paz, T. A. Harbig, D. Varga, E. Kränzle, K. Nieselt

Institute of Bioinformatics and Medical Informatics - University of Tübingen

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

CMFI Cluster of Excellence — Controlling Microbes to Fight Infections

antibiotic cellMAP

Evidente

Motivation

The evolutionary history of bacterial strains can help understand the acquisition of specific characteristics, e.g. antibiotic resistance mechanisms
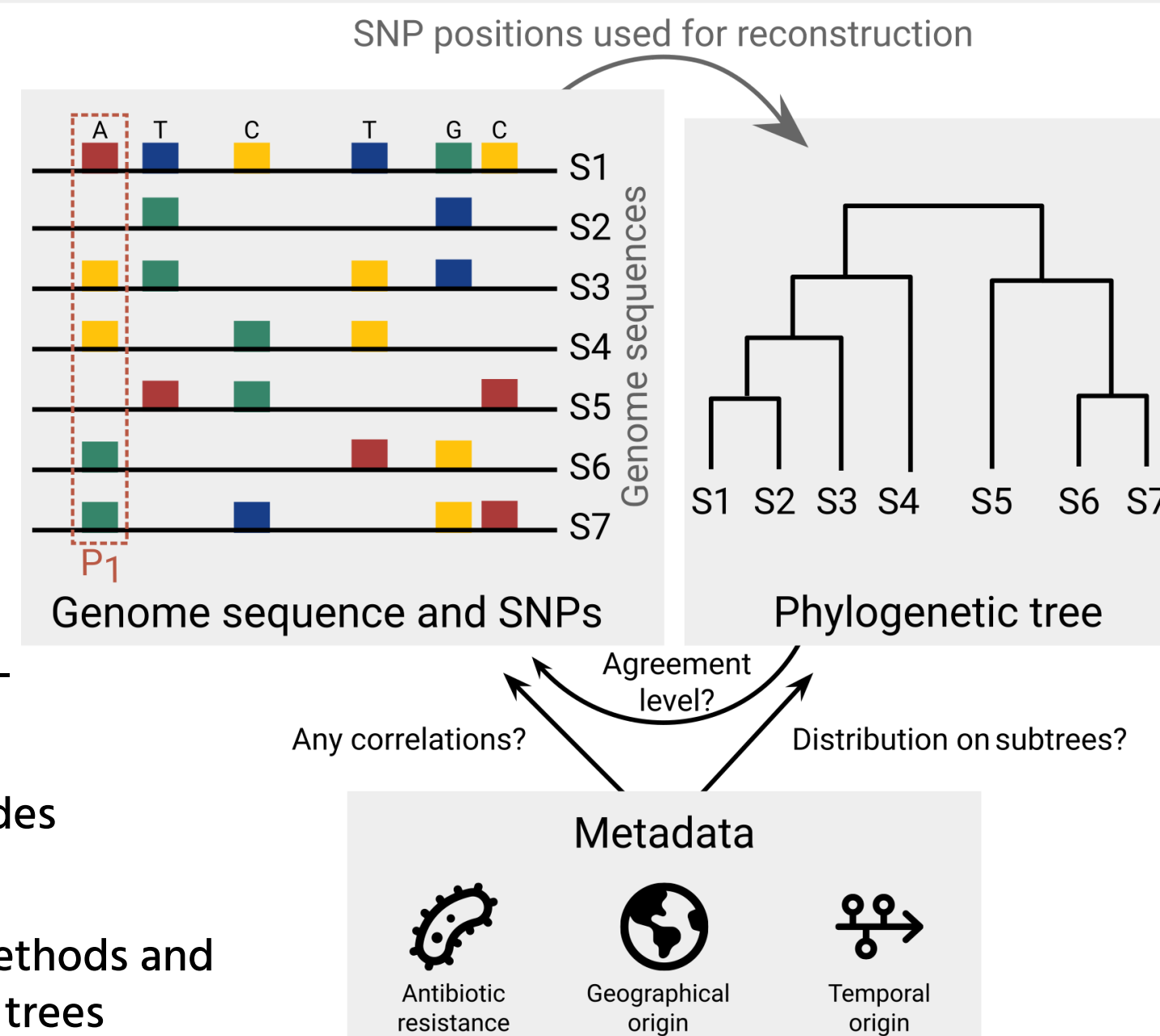
Single-point mutations in the genome (SNPs) are commonly used for the phylogenetic tree reconstruction

Reconstructed trees are rarely linked back to the underlying genomic data

Current methods lack the analytical and visual linkage between these data types – e.g. via their agreement level – on top of the metadata visualization

The recently published visual analytics tool Evidente provides linkage between these different data types [1]

Recent enhancements, such as data-driven aggregation methods and semantic zooming, improve the interaction for even larger trees

SNP positions used for reconstruction

Genome sequence and SNPs — Phylogenetic tree

Agreement level? — Any correlations? — Distribution on subtrees?

Metadata — Antibiotic resistance — Geographical origin — Temporal origin

Tasks

T1 – Improve the explainability between the phylogenetic tree and the genomic data

T2 – Identify possible patterns of SNPs or metadata within subtrees, also known as clades

T3 – Find non-clade specific patterns of SNPs and metadata across the phylogenetic tree

T4 – Allow the comparison of multiple strains

Design choices

Classification of SNPs based on their agreement with the clades of the tree: supporting and non-supporting SNPs (T1)

A dendrogram (I) aligned to two heat maps – SNP data (II) and metadata (III) – for visual linkage across data types (T2, T3)

Interacting with the nodes of the tree show the linkage between SNPs and the tree structure, based on their agreement with the tree (T2)

Filtering and aggregation techniques for a better comparison of multiple strains (T4)
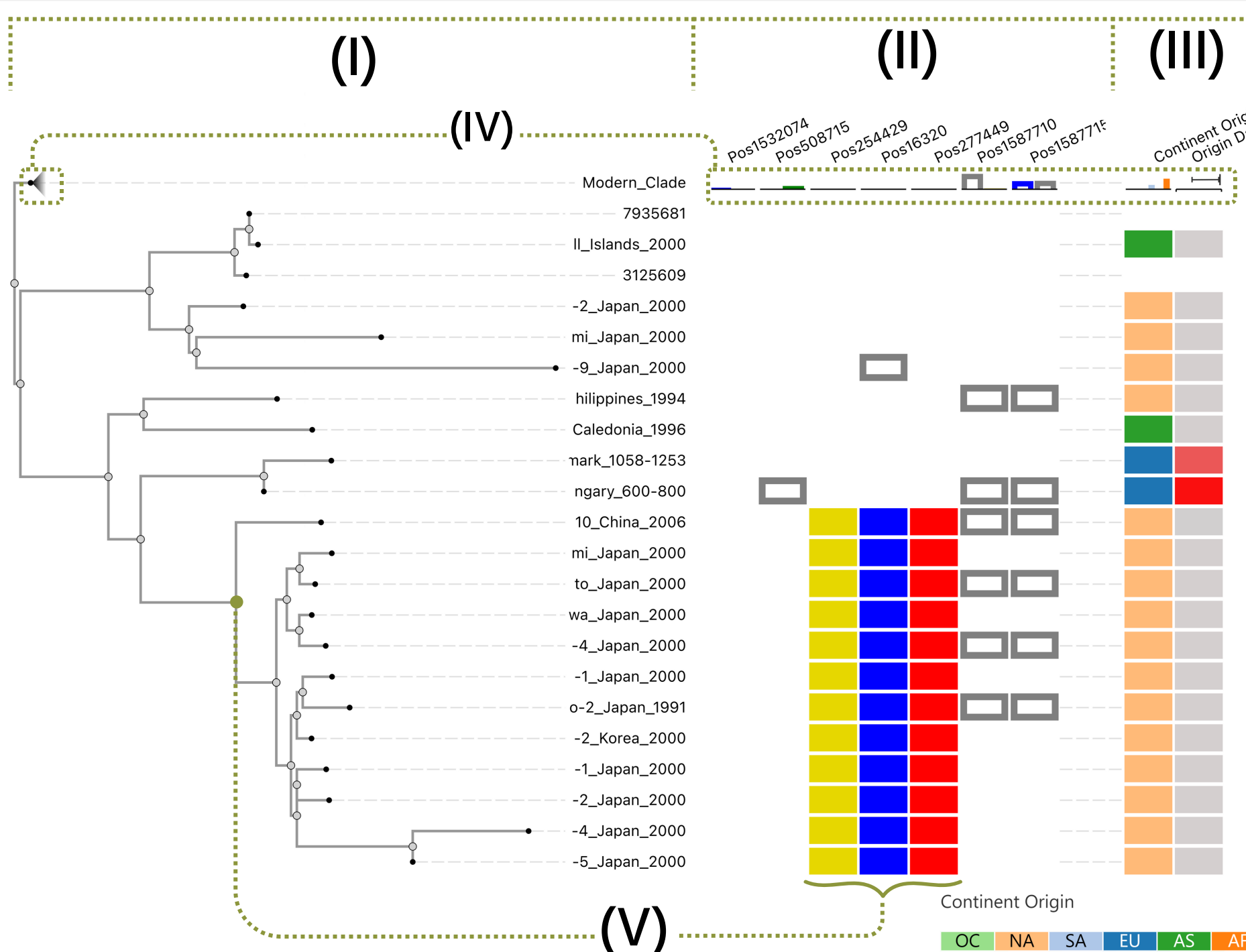
(I) Dendrogram encodes the phylogenetic tree

(II) Heat map for SNP data visualization
Each column represents a SNP position P_n in the genome. Colors encode the base of the SNP, while borders show the SNP classification.

(III) Heat map for metadata visualization

(IV) Collapsing clades summarizes data
A glyph in the tree encodes for collapsed clades, and the columns on the heat maps are aggregated based on their data type.

(V) Selecting a node shows all SNPs for the clade
For example, the three indicated SNPs are supporting the highlighted clade

Continent Origin — OC NA SA EU AS AF
Origin Date — 700 2016
SNP — A C T G N — sup. — non-sup.

Data-driven aggregation

Overview via automatic collapsing methods based on different metrics

Demo video

SNP content — Metadata — Depth of ancestor — Label selection
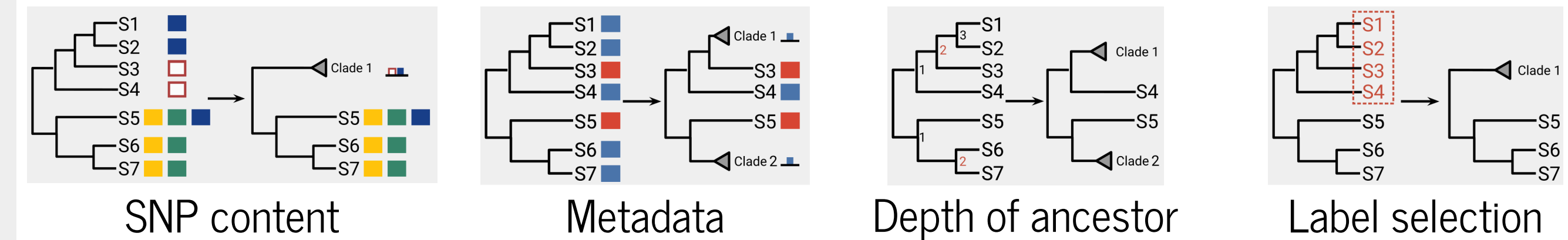
Semantic zooming

Axis-independent features improves the control of the visualization

Demo video


Geometric approach — Semantic approach
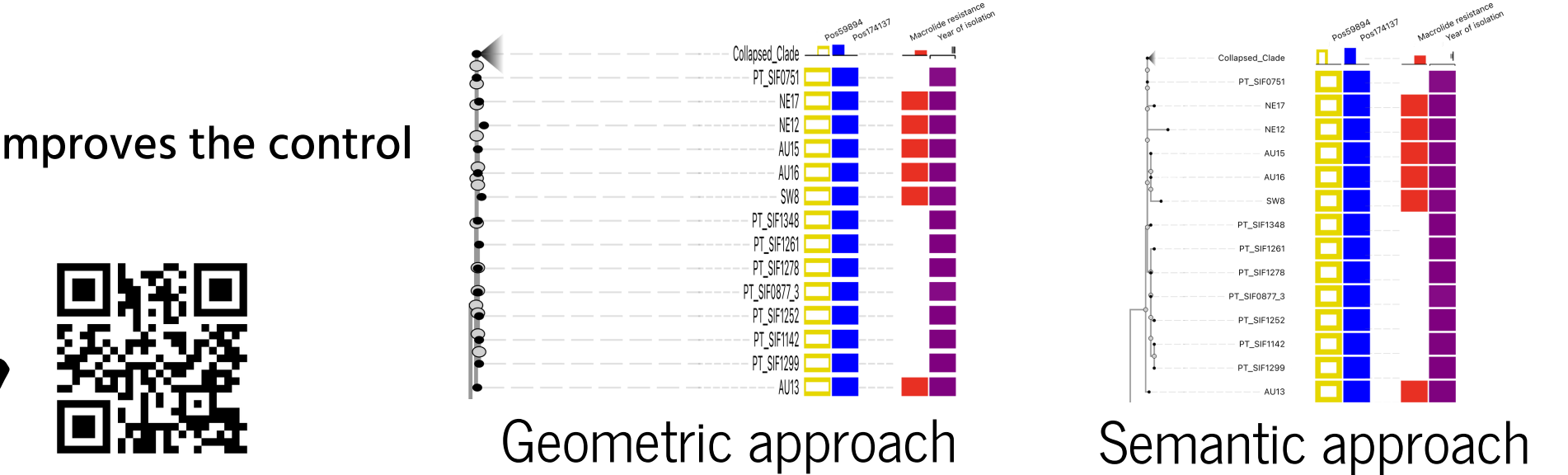
Use case on *Treponema pallidum*

Phylogenetic reconstruction of *Treponema pallidum* strains [2]

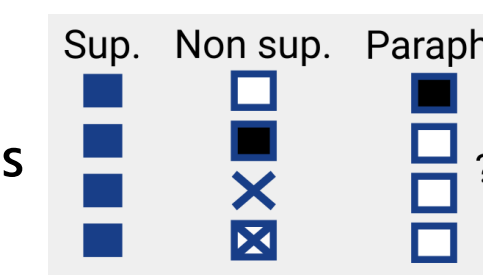Overview of the tree via aggregation based on the metadata antibiotic resistance

Interactively identify SNPs that occur in all antibiotic samples

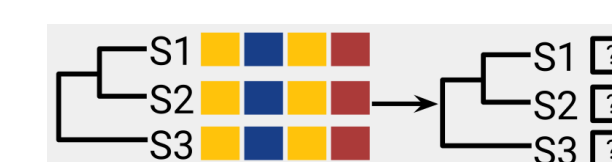Scan for the visualization process

Macrolide resistance — sensitive — resistant
SNP — A C T G N/A — sup. — non-sup.
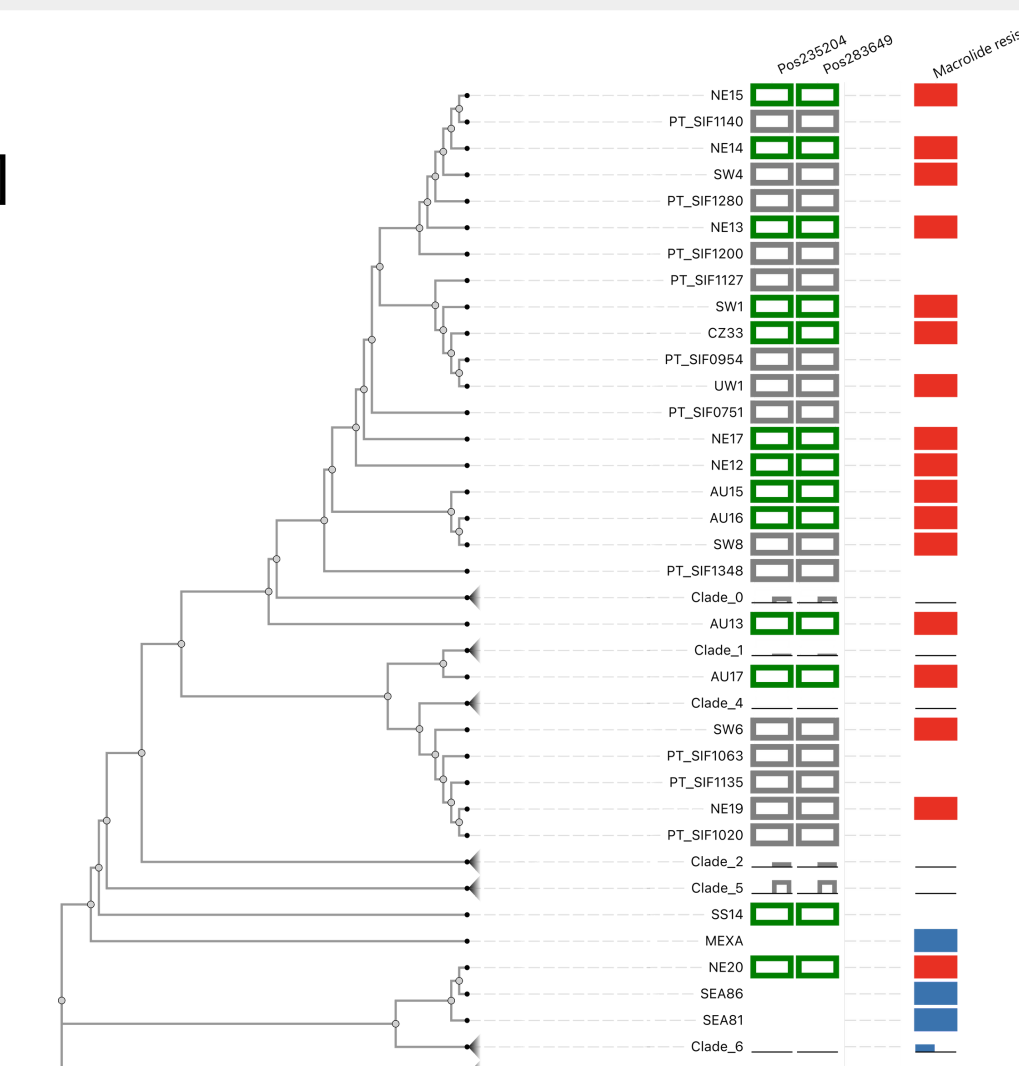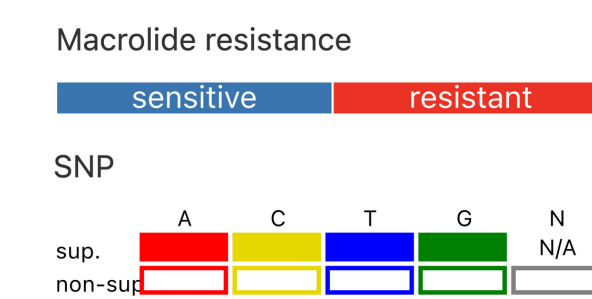
Outlook

Visualize further SNP classes
Sup. — Non sup. — Paraph.

Aggregation of many SNP columns
S1 S2 S3 → S1 S2 S3

Find more about Evidente!

tuevis.cs.uni-tuebingen.de/evidente

Funding

References

[1] Witte Paz, M., Harbig, T. A., Nieselt, K. (2022). Evidente — a visual analytics tool for data enrichment in SNP-based phylogenetic trees. Bioinformatics Advances, 2(1)

[2] Pla-Díaz, M., et al. (2022). Evolutionary processes in the emergence and recent spread of the syphilis agent, Treponema pallidum. Molecular biology and evolution, 39(1)