EVIDENTE facilitates the identification of enriched characteristics in SNP-based phylogenetic trees

Mathias Witte Paz, Alexander Seitz, Kay Nieselt

Institute for Bioinformatics and Medical Informatics
University of Tübingen
Sand 14, 72076 Tübingen

Motivation for EVIDENTE: In recent years the developments of the next-generation sequencing technologies have enabled genome resequencing projects of many individuals within one species. The genomes are often analyzed with respect to single-nucleotide polymorphisms (SNPs) or small indels. This gives the possibility of reconstructing a **phylogenetic tree** of all individuals based on the detected mutations. From such a phylogenetic tree, a common question is to **identify clade-specific SNPs** within the reconstructed phylogeny, i.e. that support the computed topology. Then one also often wishes to analyze these mutations in more detail to retrieve for example functional consequences that the SNP may have on the organism or to compute **enrichment** of certain features within the phylogenetic tree. Here, we present on-going work in developing the **visual analytics tool Evidente** for annotation and analysis of metadata in SNP-based phylogenetic trees. It enables the user to get a visual overview of distribution of SNPs across all samples as well as clade-specific SNPs within the tree. Furthermore, Evidente allows the user to run an enrichment analysis, for example for Gene Ontology (GO) annotations.

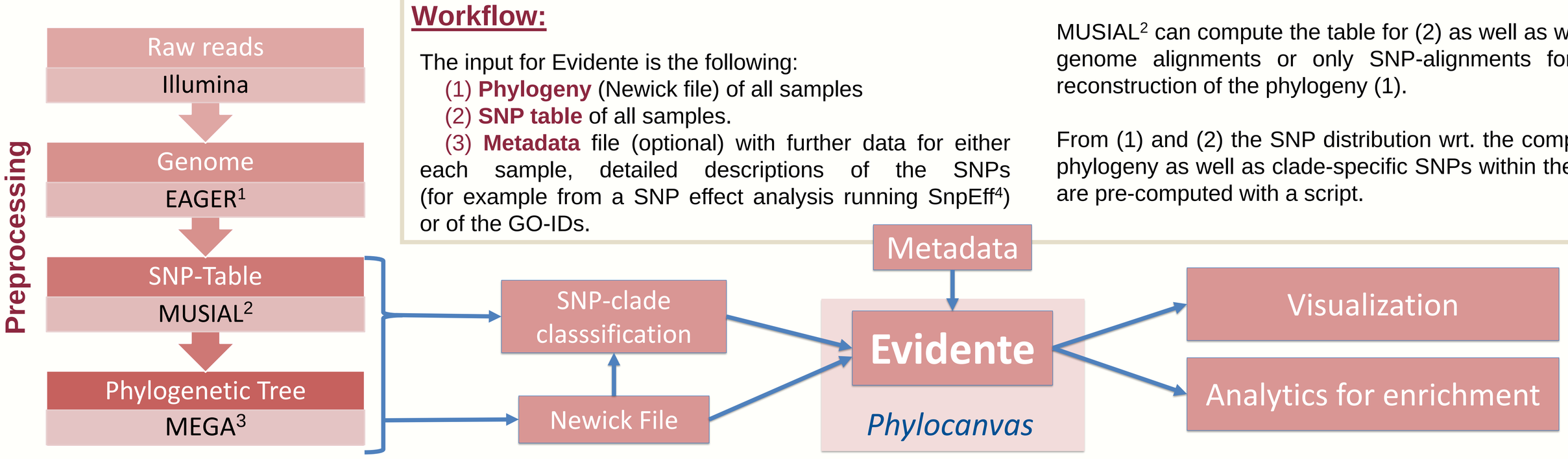
Workflow:

The input for Evidente is the following:

(1) **Phylogeny** (Newick file) of all samples
(2) **SNP table** of all samples.
(3) **Metadata** file (optional) with further data for either each sample, detailed descriptions of the SNPs (for example from a SNP effect analysis running SnpEff[4]) or of the GO-IDs.

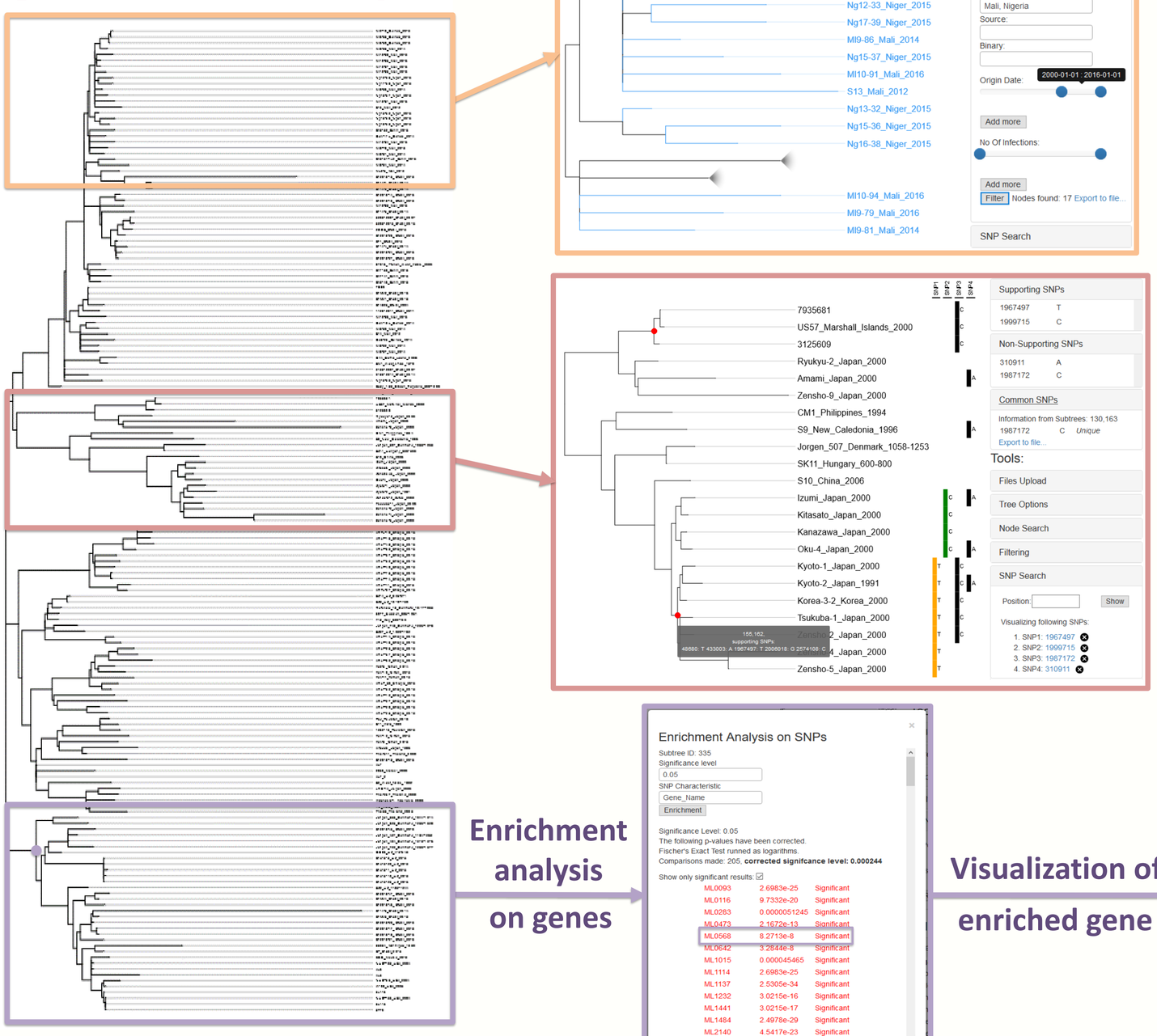
MUSIAL[2] can compute the table for (2) as well as whole-genome alignments or only SNP-alignments for the reconstruction of the phylogeny (1).

From (1) and (2) the SNP distribution wrt. the computed phylogeny as well as clade-specific SNPs within the tree are pre-computed with a script.



Use case:

***Mycobacterium leprae*[5]**

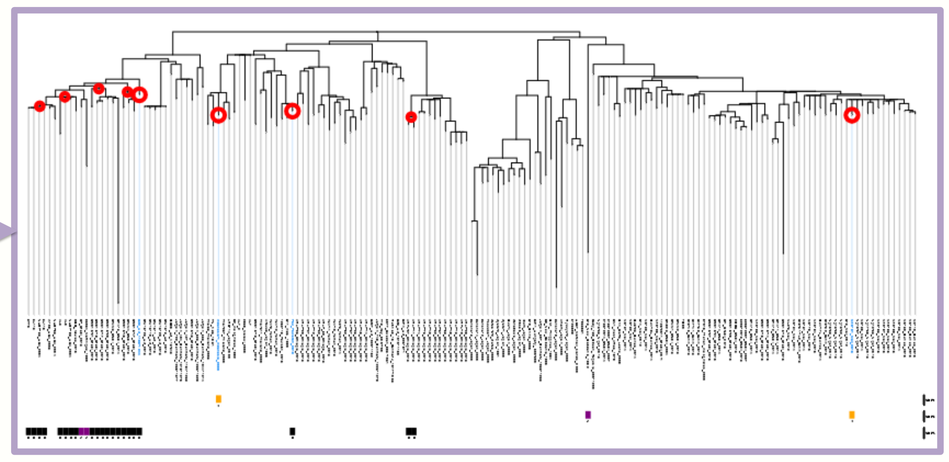


Features:

- **Filtering:**
  - Through metadata of the samples
- **SNP Visualization:**
  - Clade-specific SNPs
  - Non-supporting SNPs
  - Common SNPs between subtrees
- **Enrichment Analysis:**
  - Fisher's Exact Test
  - Enrichment of SNP characteristics
  - Enrichment of taxonomic information
  - Visualization of enriched features

Enrichment analysis on genes

Visualization of enriched gene



References

1) Peltzer, A. et al. EAGER: efficient ancient genome reconstruction. *Genome Biol* 17:60 (2016).
2) https://github.com/Integrative-Transcriptomics/MUSIAL
3) Kumar, S. et al. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets, *Molecular Biology and Evolution*, 33: 1870-1874 (2016).
4) Cingolani, P. et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly* 6:80-92 (2012).
5) Schuenemann, VJ. et al. Ancient genomes reveal a high diversity of *Mycobacterium leprae* in medieval Europe. *PLOS Pathogens* 4: e1006997(2018).

Availability:

https://lambda.informatik.uni-tuebingen.de/gitlab/paz/evidente

Poster download!