Whole genome SNPs comparison

Valeria Michelacci

Bioinformatics training, June 2018

Reference-based wgSNPs typing

- **Alignment** to a reference sequence
- Compiling of a **variant call format file** per strain
- Compiling of a **distance matrix**
- **Phylogenetic tree** built on the distance matrix

Tools available for download – possibility to build your own pipeline

CGE webserver hosted by DTU offers easy to use pipelines

- **NDtree**
- **CSI phylogeny**

DTU

Center for Genomic Epidemiology

Ref-based wgSNPs/1: NDtree

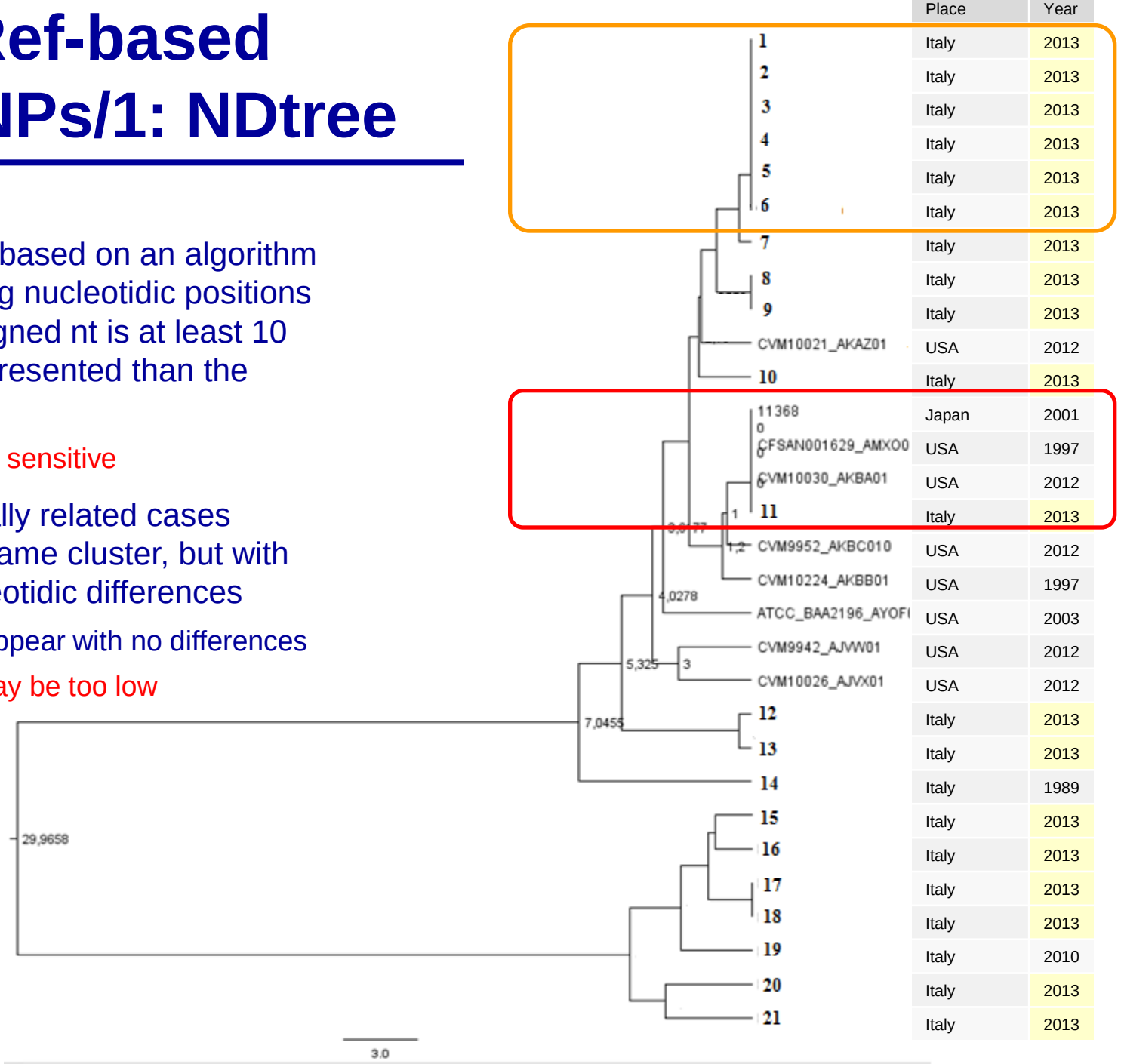
SNPs analysis based on an algorithm only considering nucleotidic positions where the assigned nt is at least 10 times more represented than the other three

More robust, less sensitive

Epidemiologically related cases appear in the same cluster, but with no visible nucleotidic differences

Very far strains appear with no differences

The sensitivity may be too low

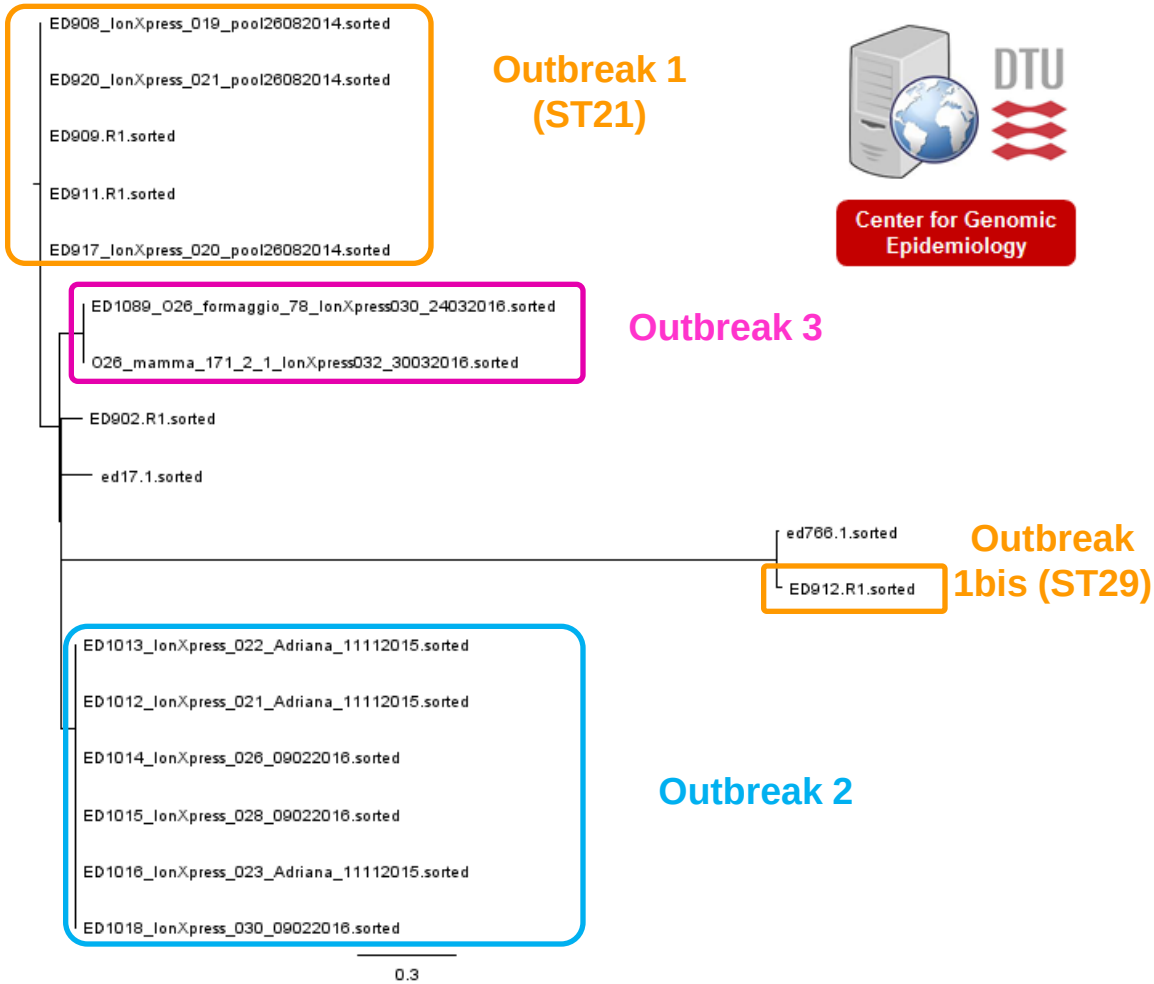


Place	Year
Italy	2013
Italy	2013
Italy	2013
Italy	2013
Italy	2013
Italy	2013
Italy	2013
Italy	2013
Italy	2013
USA	2012
Italy	2013
Japan	2001
USA	1997
USA	2012
Italy	2013
USA	2012
USA	1997
USA	2003
USA	2012
USA	2012
Italy	2013
Italy	2013
Italy	1989
Italy	2013
Italy	2013
Italy	2013
Italy	2013
Italy	2010
Italy	2013
Italy	2013

DTU

Center for Genomic Epidemiology

Ref-based wgSNPs/2: CSI phylogeny



DTU

Center for Genomic Epidemiology

CSI phylogeny

- It only computes positions having a good quality score in all the strains tested
- Only feed good sequences to avoid reducing the amount of computed positions!
- It accepts reads and/or contigs, but at least some samples need to be uploaded as reads to allow the computation of SNPs quality

Epidemiological clusters identified, but no difference among strains

Ref-based wgSNPs/3: PHEnix

Phoenix

latest

Search docs

Introduction
Prerequisites
Filters
Annotators
Calling SNPs
Converting to FASTA
API Reference
Scripts
Galaxy

Read the Docs v: latest

Docs »
Public Health England SNP calling Pipeline (PHEnix)
Edit on GitHub

Public Health England SNP calling Pipeline (PHEnix)

This documentation is designed to give an overview as well as detailed API reference for Public Health England's single nucleotide polymorphism calling pipeline. Currently the pipeline does not provide stand alone means of calling SNPs, instead it interfaces with other published tool e.g. BWA and samtools.

Contents:

http://phenix.readthedocs.io/en/latest/index.html

Ref-based wgSNPs/3: PHEnix

Overview

This code was designed to allow users to input fastq files and a reference sequence and perform:

- Reference mapping
- VCF generation
- VCF filtering
- FASTA sequence of SNPs

Tools available for download

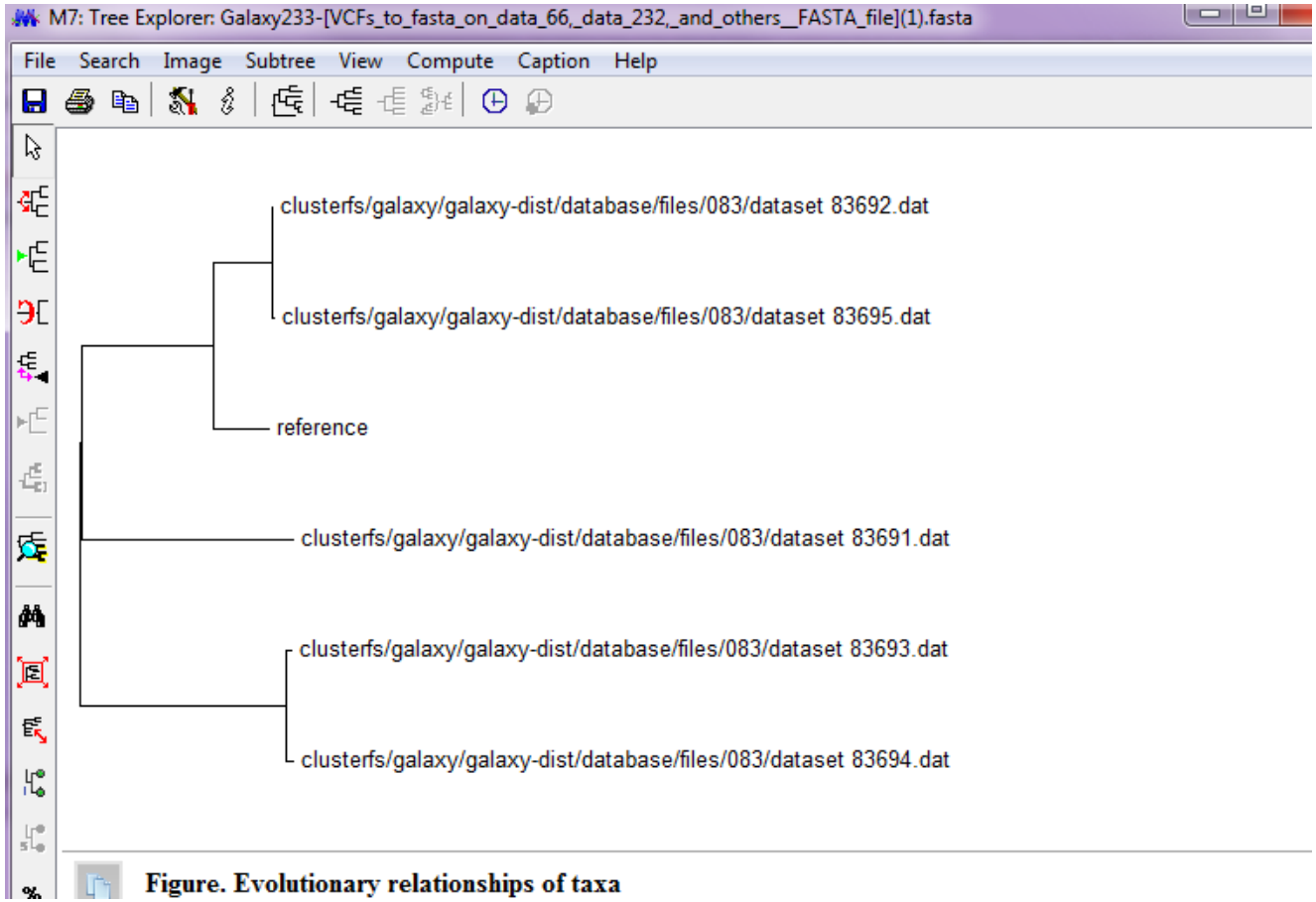
Executable through command line

Available for installation on Galaxy

Up and running on ARIES

Ref-based wgSNPs/3: PHEnix

The output is a multifasta to be used to infer phylogeny after clustering through an appropriate software



FDA SNPs pipeline

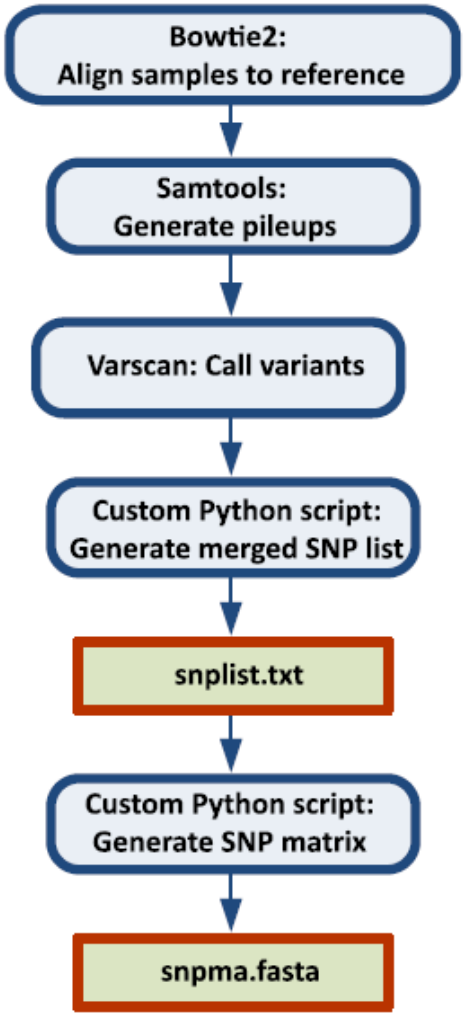


Figure 1 **Steps in the SNP Pipeline.** Rounded blue outlined boxes are analysis steps and squared red outlined boxes are files produced by the pipeline.

The output is a multifasta to be used to infer phylogeny after clustering through an appropriate software

Reference-free wgSNPs typing

- **ksnp3** looks for SNPs in central positions of k-mers

  The optimal length of the kmer is computed for every batch of test sequences

- It accepts **fasta** files

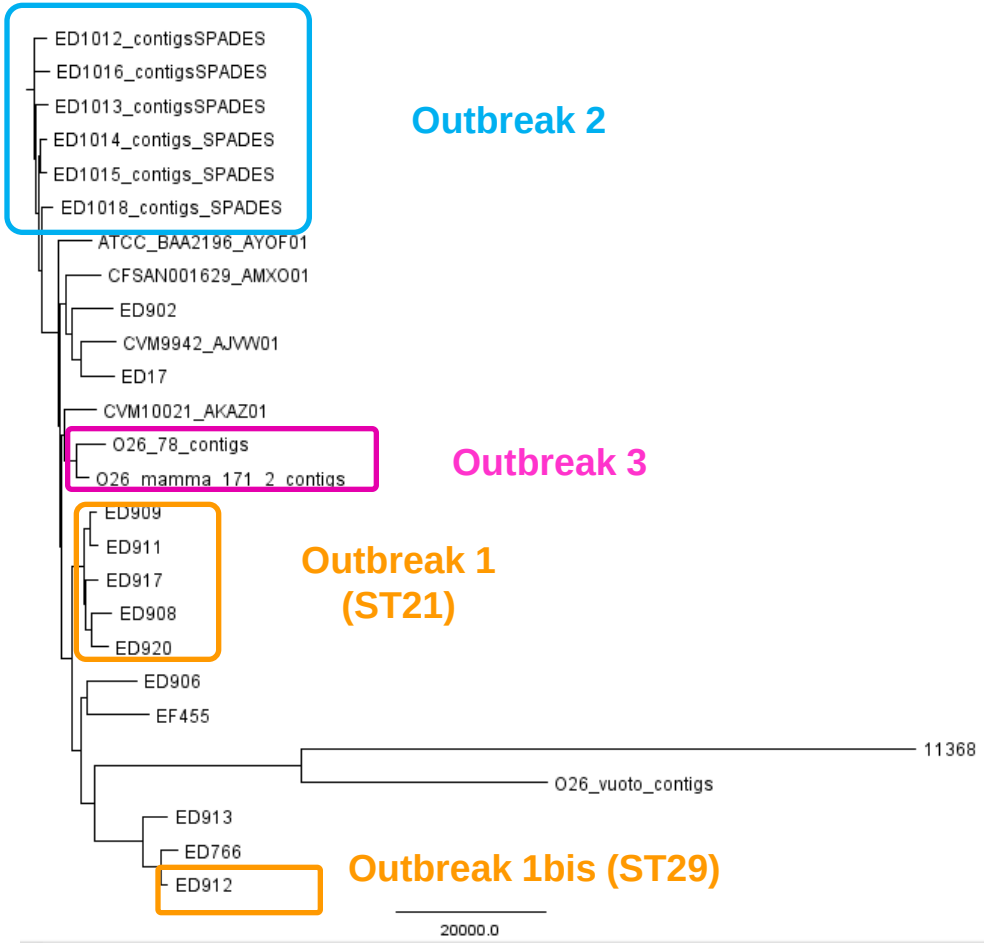
- **Different clustering algorithms** available

Available for download as a tool package operated via command line

Available on ARIES (www.iss.it/site/aries)

Ref-free wgSNPs typing: ksnp3

ksnp3 - ARIES



Galaxy / ARIES - ISS
Istituto Superiore di Sanita'
www.iss.it/site/aries

- **Epidemiological clusters correctly identified**
- **Intra-cluster discrimination**

ksnp3 - ARIES

E. coli typing - phylogeny

MakeKSNP3infile

MakeFasta

Kchooser

It identifies an optimum value of k for your specific data set

kSNP

Flag the options for estimating Neighbor Joining and maximum likelihood trees