Whole genome SNPs comparison: SNPtree, NDtree, CSI Phylogeny and kmer-based analysis

Valeria Michelacci

NGS course, June 2016

Istituto Superiore di Sanità, Dip. Sanità Pubblica Veterinaria e Sicurezza Alimentare,
Laboratorio Europeo e Nazionale di Riferimento per *E. coli*

Webserver-based free pipelines available

- **NCBI pipeline**
- **FDA GenomeTrakr**
- **CGE/DTU Batch upload and analysis**

Assembly **(Annotation)** **Clustering based on SNPs analysis**

Storage at International Nucleotide Sequence Database Collaboration (INSDC)
NCBI – EMBL – DDBJ

Direct or after embargo period

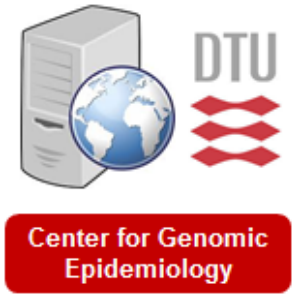
Reference-based wgSNPs typing

- **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

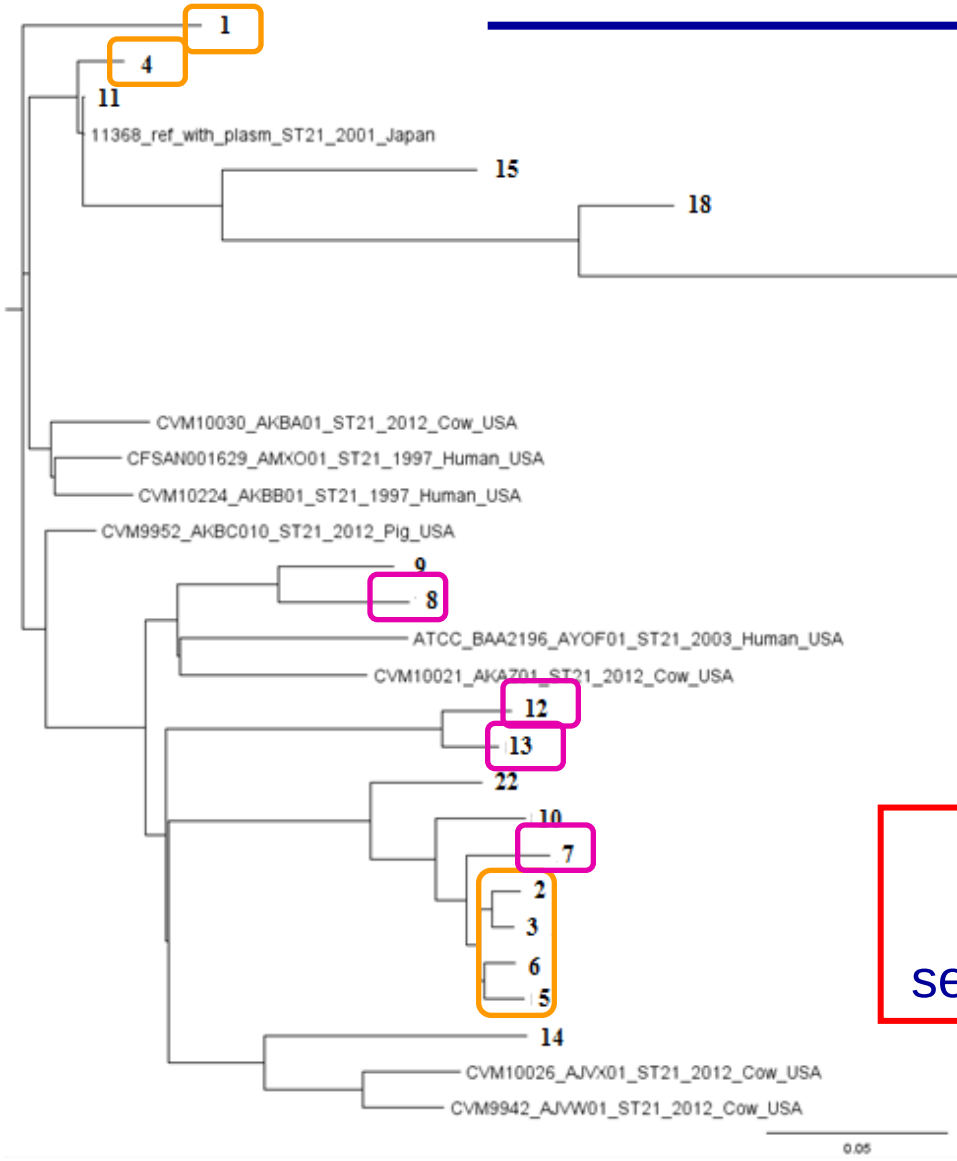
Tools available for download – possibility to build your own pipeline

CGE webserver hosted by DTU offers easy to use pipelines

- **SNPs tree**
- **NDtree**
- **CSI phylogeny**



Ref-based wgSNPs/1: SNPs tree



Leekitcharoenphon P,
BMC Genomics 2012

DTU
Center for Genomic Epidemiology

= PFGE correlated isolates, Italy, August 2013

= not correlated isolates, Italy, August 2013

Analysis of all the detected nucleotidic differences based on the comparison with a reference sequence

Epidemiologically related cases appear very different. Such a high sensitivity increases the risk for errors

Discontinued on DTU CGE webserver

Ref-based wgSNPs/2: NDtree

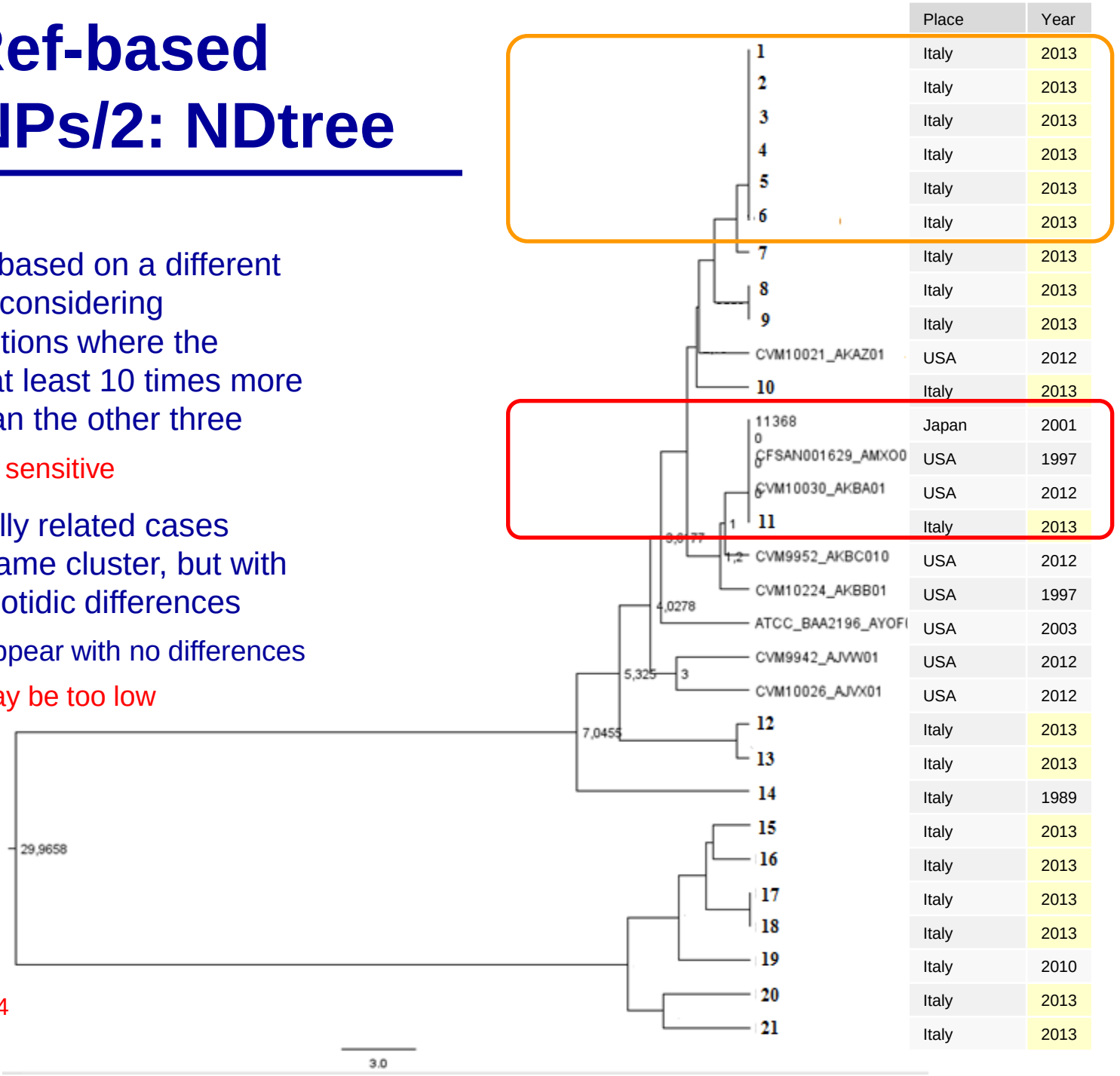
SNPs analysis based on a different algorithm: only considering nucleotidic positions where the assigned nt is at least 10 times more represented than the other three

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

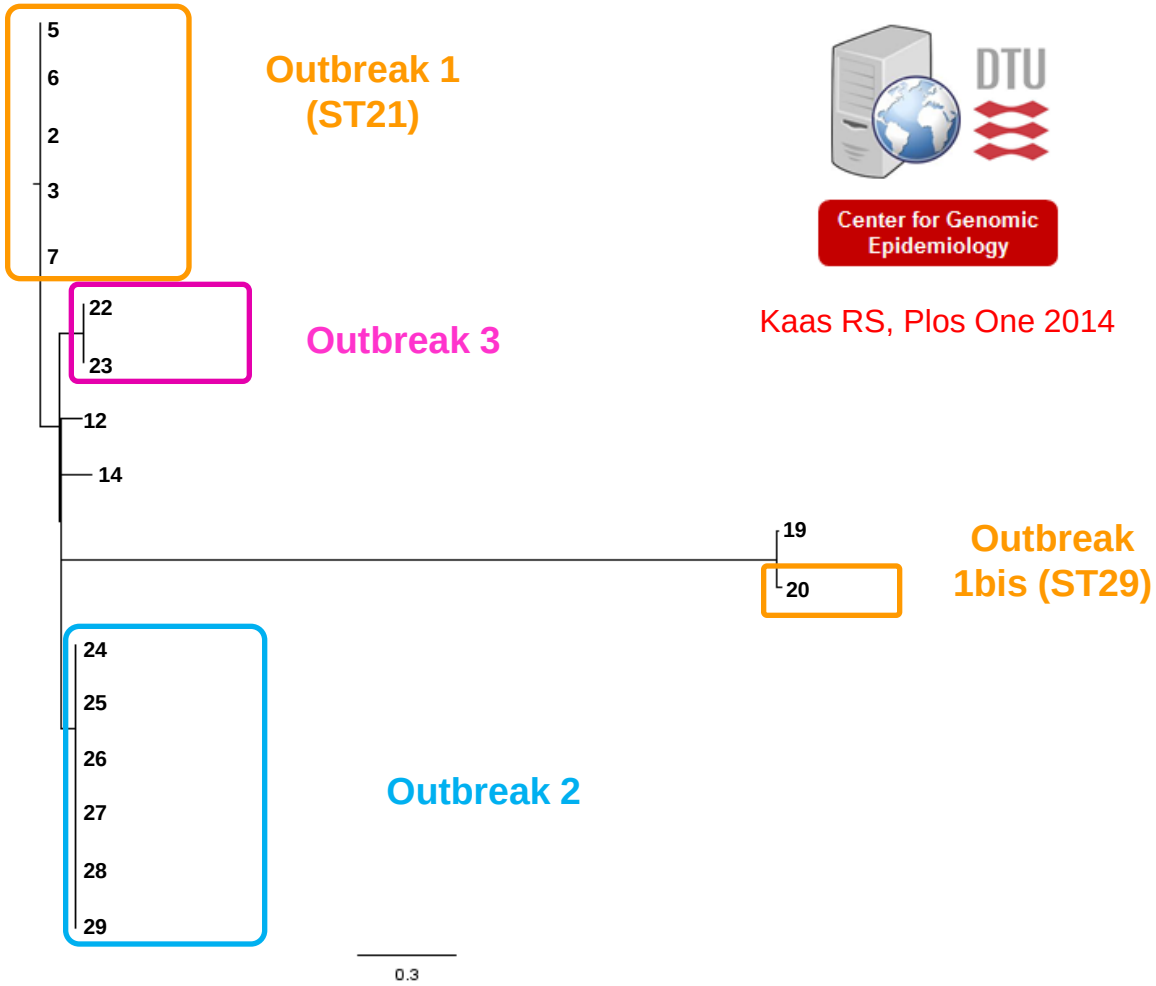
The sensitivity may be too low



Center for Genomic Epidemiology

Joensen KG, JCM 2014

Ref-based wgSNPs/3: CSI phylogeny



Center for Genomic Epidemiology

Kaas RS, Plos One 2014

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

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

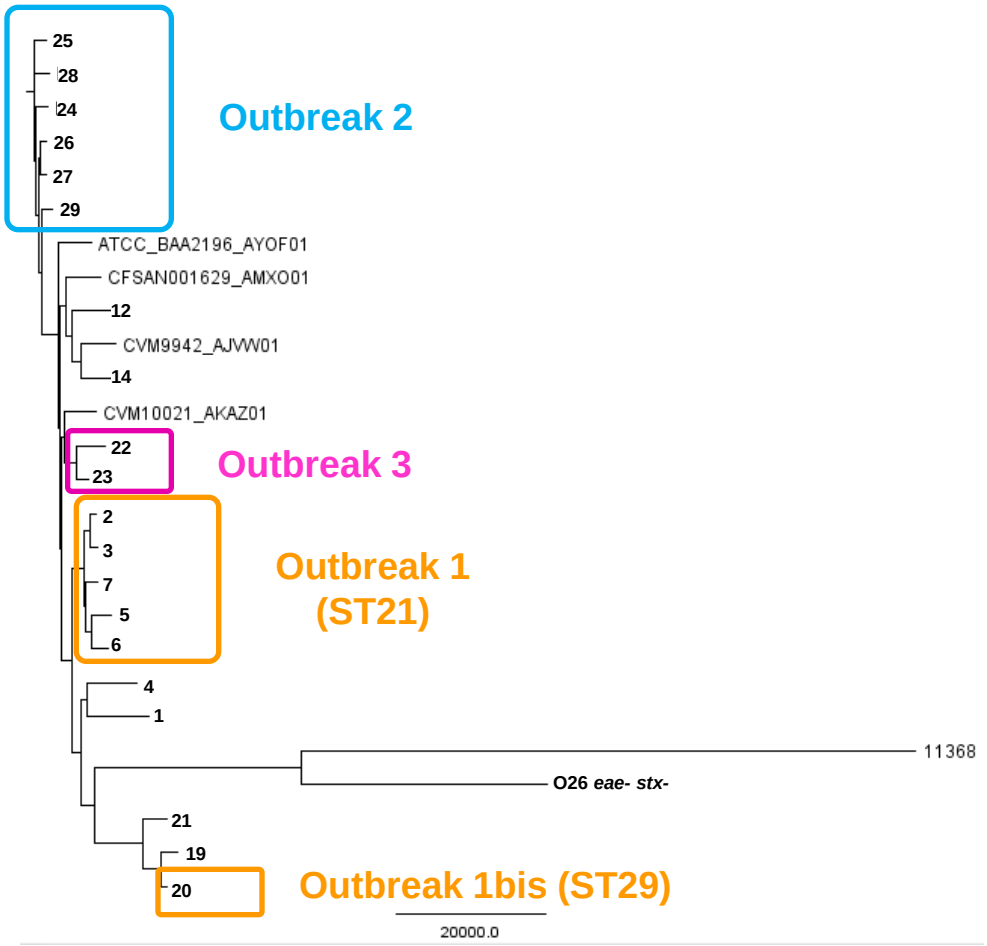
- **Different clustering algorithms** available

Available for download as a tool package operated via command line

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

Ref-free wgSNPs typing: ksnp3 (through ARIES)

Gardner SN, Bioinformatics 2015



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

ksnp3 – ARIES: how does it work?

E. coli typing - phylogeny