Typing in the NGS era: The way forward!

Valeria Michelacci

NGS course, June 2015

Typing from sequence data

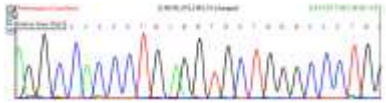
- NGS-derived **conventional Multi Locus Sequence Typing** (University of Warwick, 7 housekeeping genes)
- **SNPs analysis**: whole genome comparison of single nucleotidic polymorphisms
- Whole Genome MLST (**wgMLST**), Core Genome MLST (**cgMLST**) and more

Still widely open for development!

7-genes MLST

Conventional Sanger sequencing

PCR, sequencing, electropherograms analysis



Uploading sequences on a webserver to obtain the corresponding alleles and STs

THE UNIVERSITY OF WARWICK

NGS-derived

Direct upload of WGS contigs on a webserver
(e.g. ARIES or CGE)

Alleles are directly retrieved through blastn comparison with pre-installed database of alleles from University of Warwick with pre-compiled pipelines

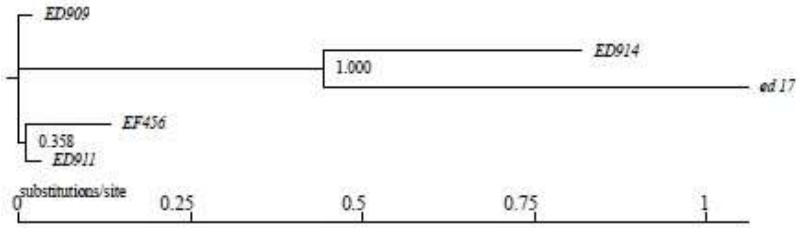
The way forward: amplifying the 7 genes, pooling, shearing the pool.

About 7x800bp = 5600 bp per sample

Possibility to barcode the pools and perform simultaneous MLST of wide panels of strains via NGS

Single Nucletidic Polymorphisms concept

- **Multiple alignment** of the WGS of the test strains
- Compiling of a **variant call format file** per strain, containing the information about each SNP identified
- Compiling of a **distance matrix**
- **Phylogenetic tree** built on the distance matrix



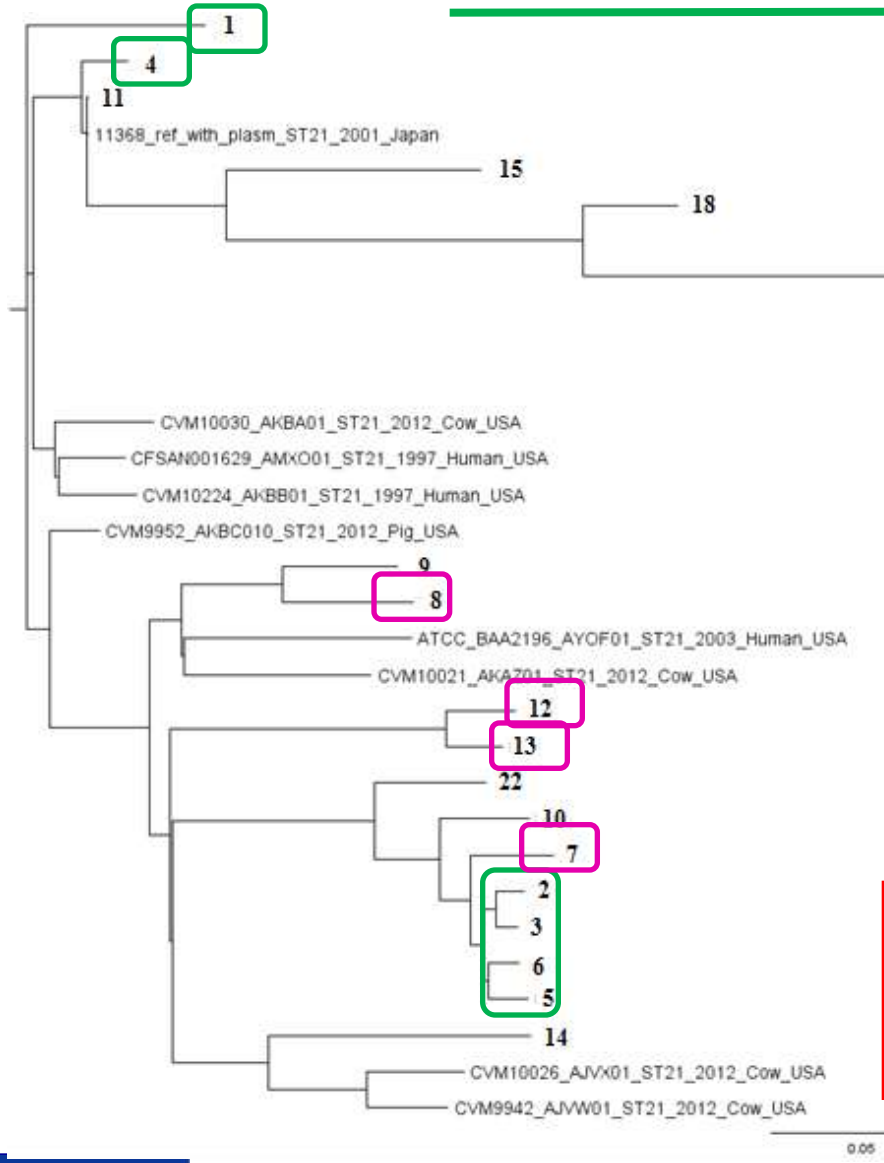
High sensitivity!

Still in implementation:
need to only evaluate true differences

Parameters:

- Minimum coverage
- Minimum SNP quality
- Minimum read mapping quality
- Minimum distance between SNPs
- Minimum distance to end of the sequence

SNPs tree



= 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

NDtree

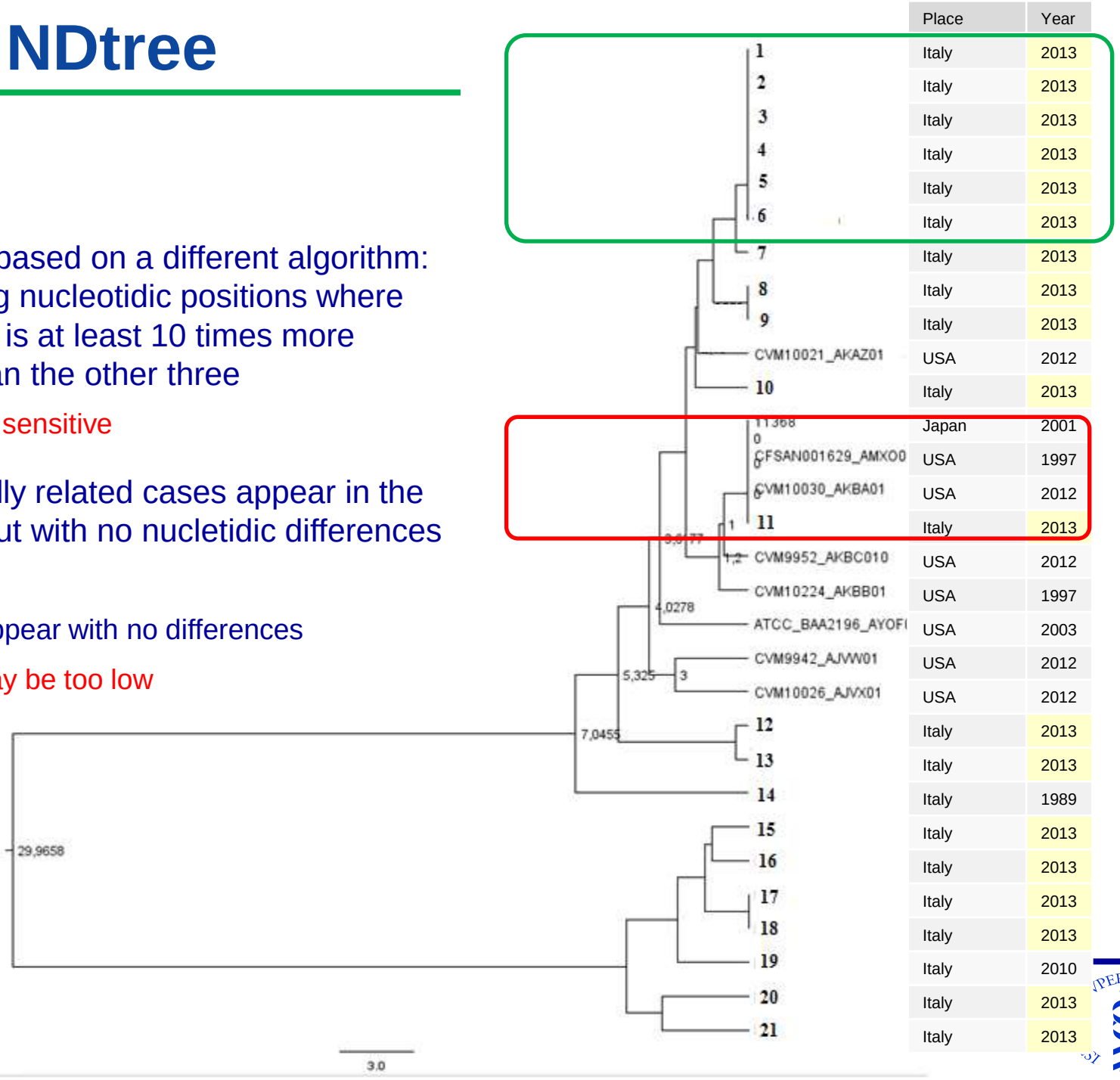
SNPs analysis based on a different algorithm: only considering nucleotidic positions where the assigned nt is at least 10 times more represented than the other three

More robust, less sensitive

Epidemiologically related cases appear in the same cluster, but with no nucletidic differences at all

Very far strains appear with no differences

The sensitivity may be too low



Whole Genome MLST

Need for biologically consistent bioinformatic tools for NGS data analysis

Whole genome-based MLST (wgMLST):
Analysis of the SNPs in a wide panel of genes

Allelic profiling of the strains based on wide panels of genes

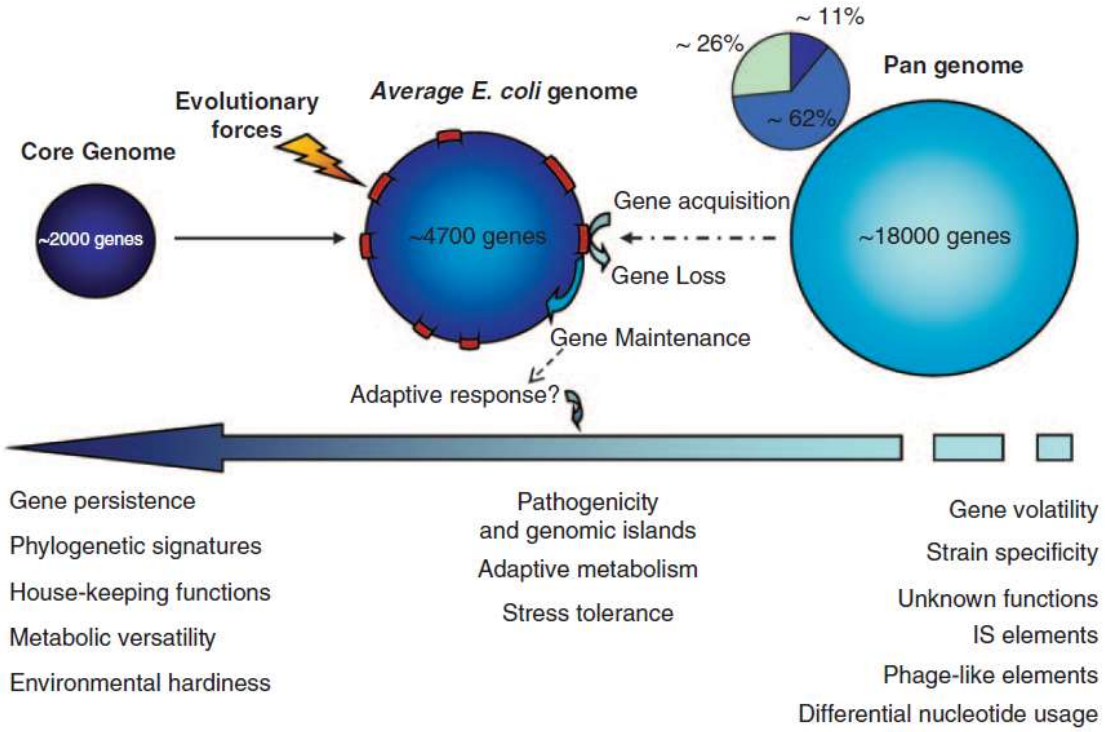
Core Genome MLST (cg-MLST): MLST based on all the core genes of the species

Need for **reference databases** of alleles sequences and codification in **cg-sequence types** (cg-ST, combination of alleles) and **cg-clonal complexes** (cg-CC, grouping cg-STs sharing a proportion or subset of alleles of the core genes)

The *E. coli* pangenome

Genomic plasticity

Huge pan-genome



Van Elsas J.D. et al., 2011

Pangenome **Whole genome** **Core genome** **Accessory genome**

Housekeeping

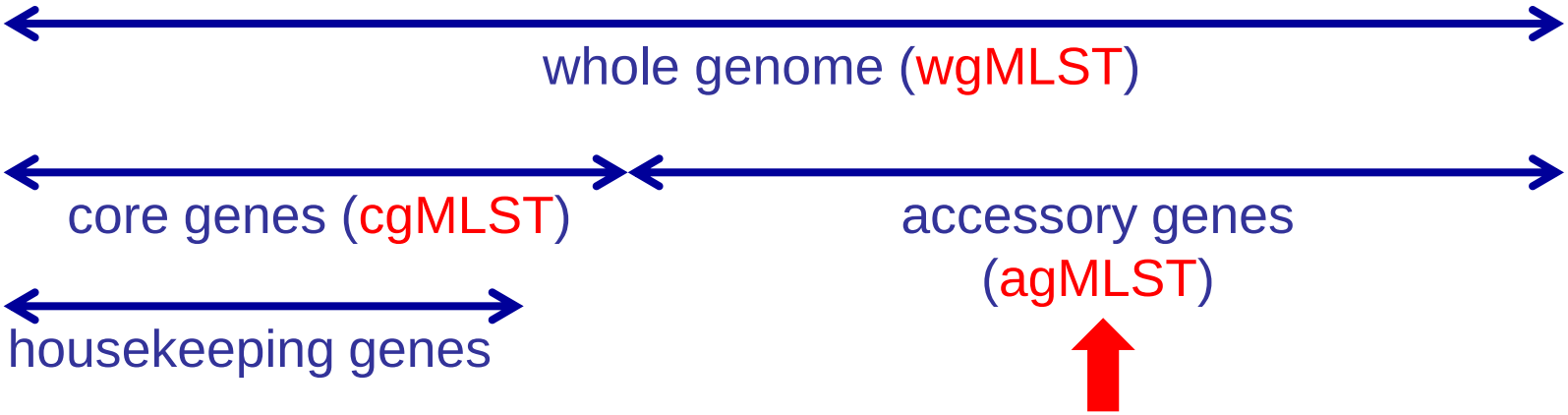
Applying MLST to *E. coli*

Conventional MLST

7 housekeeping genes

Low sensitivity	Good for phylogenetic analysis
High robustness	Not good enough for outbreak investigation

MLST from WGS data



Proposed nomenclature

wgMLST High sensitivity; need for a threshold?

cgMLST Lower sensitivity; is it enough for *E. coli* strain identification?

agMLST Intermediate sensitivity (low computational requirements)
High significance for *E. coli*
It could be good for *E. coli* strain identification

The result of MLST typing could take in consideration all these schemes, providing **different levels** of characterization of the strain

Scheme	Alleles		ST
wgMLST	A-D-F-C-A-…-B	Unique signature of the strain	
cgMLST	D-C-A-E	Signature of the core genes	cg-ST
agMLST	A-F-B	Signature of the accessory genes	ag-ST

Propose for communication

A complete signature of the genetic content of the strain structured in levels

Example of results:

	wgMLST signature	cg-ST	ag-ST
Strain X	A-D-F-C-A-…-B	cg-ST170	ag-ST211

For a deepest identification, **the schemes could be subdivided in subsets**: e.g. cg-ST divided in **housekeeping** and nonhousekeeping-ST and ag-ST divided in **PAIs-ST**, **plasmids-ST**...

Subsets of accessory genes could be used also for **risk assessment** and for guiding clinical menagement of the infections

The accessory genome of *E. coli*

Study on the three major PAIs of STEC in 730 strains

- **LEE** 38/43 ORFs
- **OI-122** 12/16 ORFs
- **OI-57** 41/117 ORFs

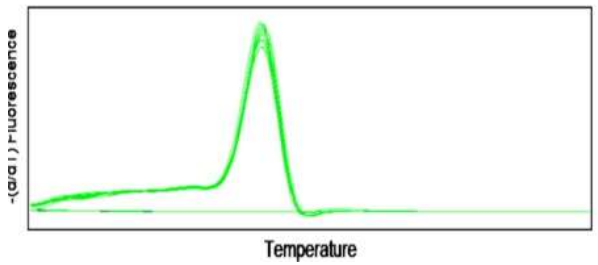
TOTAL **91/172 ORFs tested**

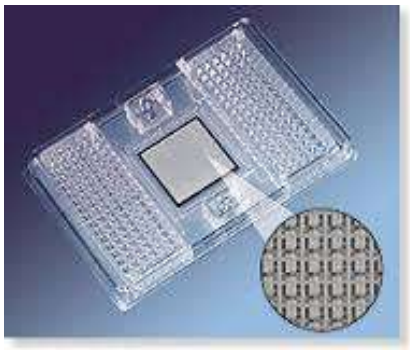
Allelic profiling of the STEC accessory genome could provide a mean to investigate the evolution of the different groups of pathogenic *E. coli*.

Valeria Michelacci[1], Massimiliano Orsini[2], Arnold Knijn[1], Sabine Delannoy[3], Patrick Fach[3], Alfredo Caprioli[1], Stefano Morabito[1]

vbs.psu.edu/research/centers/ecoli/e-coli-workshop

Analysis of the Tm of PCR amplified products





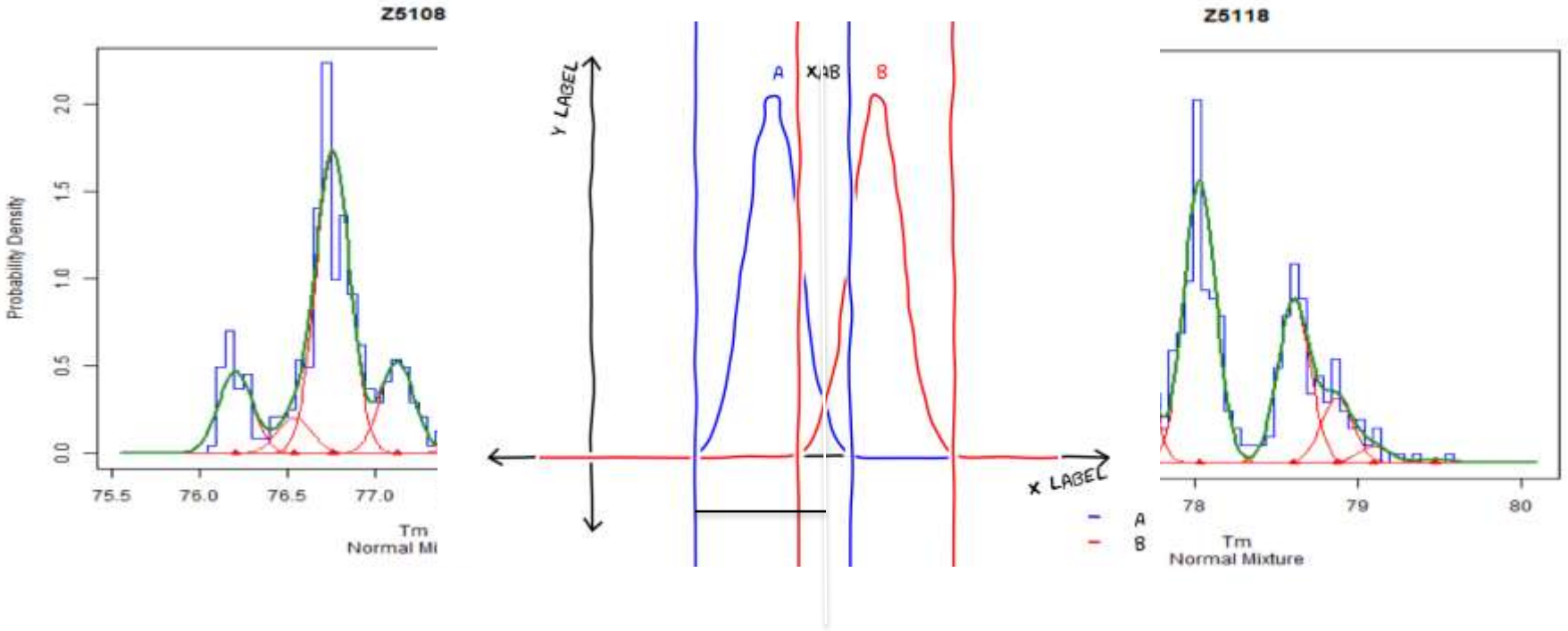
HReVAP: Automatic BIN assigment

Tm

O108:H9	DG233/8	30	73,7	77,26	78,65	80,78	83,94	76,2	78,04	77,9
O108:H9	DG239/5	31	73,78	77,35	78,74	80,87	84,05	76,3	78,13	78,01
O108:H9	DG258/1	32	73,81	77,39	78,77	80,9	84,1	76,33	78,18	78,04
O108:H9	DG314/6	33	73,81	77,38	78,77	80,89	84,09	76,34	78,17	78,04

Alleles

O108:H9	DG233/8	Z	B	B	B	C	F	Z	B	B	C	Z	Z
O108:H9	DG239/5	Z	B	B	B	C	F	Z	B	B	C	Z	Z
O108:H9	DG258/1	Z	B	B	B	C	F	Z	B	B	C	Z	Z
O108:H9	DG314/6	Z	B	B	B	C	F	Z	B	B	C	Z	Z

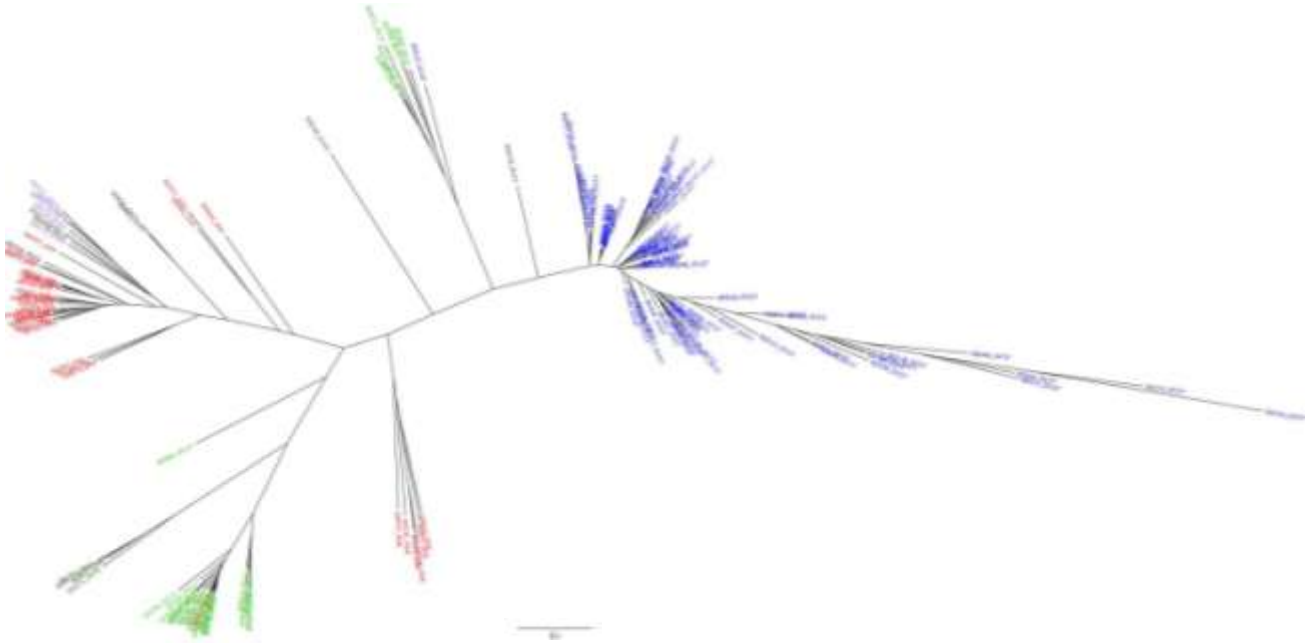


Typing *E. coli* with HReVAP

- 2 to 9 different alleles detected in each of the 91 genes tested (mean = 4.7)
- Total 435 alleles detected in the 91 genes tested (mean = 4.7) ->impressive number of combinations

The combination of alleles represents a significative signature of the tested strain

Clustering by HReVAP



HReVAP: cross-platform and cross-generation

- HReVAP allows following the **evolution of the MGEs** by using subpanels of PAIs for the analysis
- Analysis of accessory genes proved valuable for *E. coli* **typing** (identification of sub-populations of VTEC even within serogroups) → **SURVEILLANCE?**
- It could be applied to the accessory genome of **other pathogens**

Already developed in house for ARIES, soon open and running

RT-PCR

Sanger

NGS

Possibility to translate the allele calling to use sequence data!

HReVAP could be expanded to the whole accessory genome

Designation of alleles basing on an established panel of Tm intervals avoids the problems due to base calling problems!

Conclusions

In order to use NGS data for surveillance **Standard Operative Procedures** are needed for:

- Quality check
- Filtering
- Assembling
- NGS-based typing

Need for **reference databases for MLST schemes**

HReVAP-based tools virtually do not need reference databases, but need for curation for new alleles detected

Central repository could receive only allelic profiles or Sequence Types identifiers and the log files of the analysis