WHOLE GENOME SEQUENCING @ ISS

EU RL for *E. coli*

Valeria Michelacci

Copenhagen, January 29th 2015

Availability of Samples and WGS technology

Strains collection ≅ 2000 pathogenic *E. coli* strains

Clinical samples ≅ 150-200 specimens analysed/year (mainly BD, HC and HUS)

Sequencing platforms @ ISS:

2 x ION PGM

1 x MiSeq

1 x 454

Several ABI 3130 (Sanger)

Sequencing throughput and IT

Data production

TODAY

70 strains fully sequenced

TOMORROW (projection to be achieved in the next two years)

200 strains /year + 20 metagenomics samples

Data storage

TorrentServer, Central servers, firewalls, intrusion prevention systems, automated back-up services, storage units

TODAY

15 Tb dedicated to sequence storage (backup and mirroring)

TOMORROW (projection to be achieved in the next two years)

60 Tb (local)

Data analysis: Locally running softwares

- *de novo* assembly
- Alignment of sequences, production of VCF files, production of dendrograms
- MLST
- Search for interesting genes

USER-FRIENDLY INTERFACE, Slow processing, RAM needed

- *de novo* assembly
- Search for interesting genes
- Alignment of sequences, production of VCF files

BUILT IN THE ION TORRENT TECHNOLOGY PACKAGE

Data analysis: web servers

- Species identification
- *de novo* assembly tools
- VirulenceFinder
- ResFinder
- MLST
- SNPs tree and newly deleveloped NGS-driven philogenetic tools

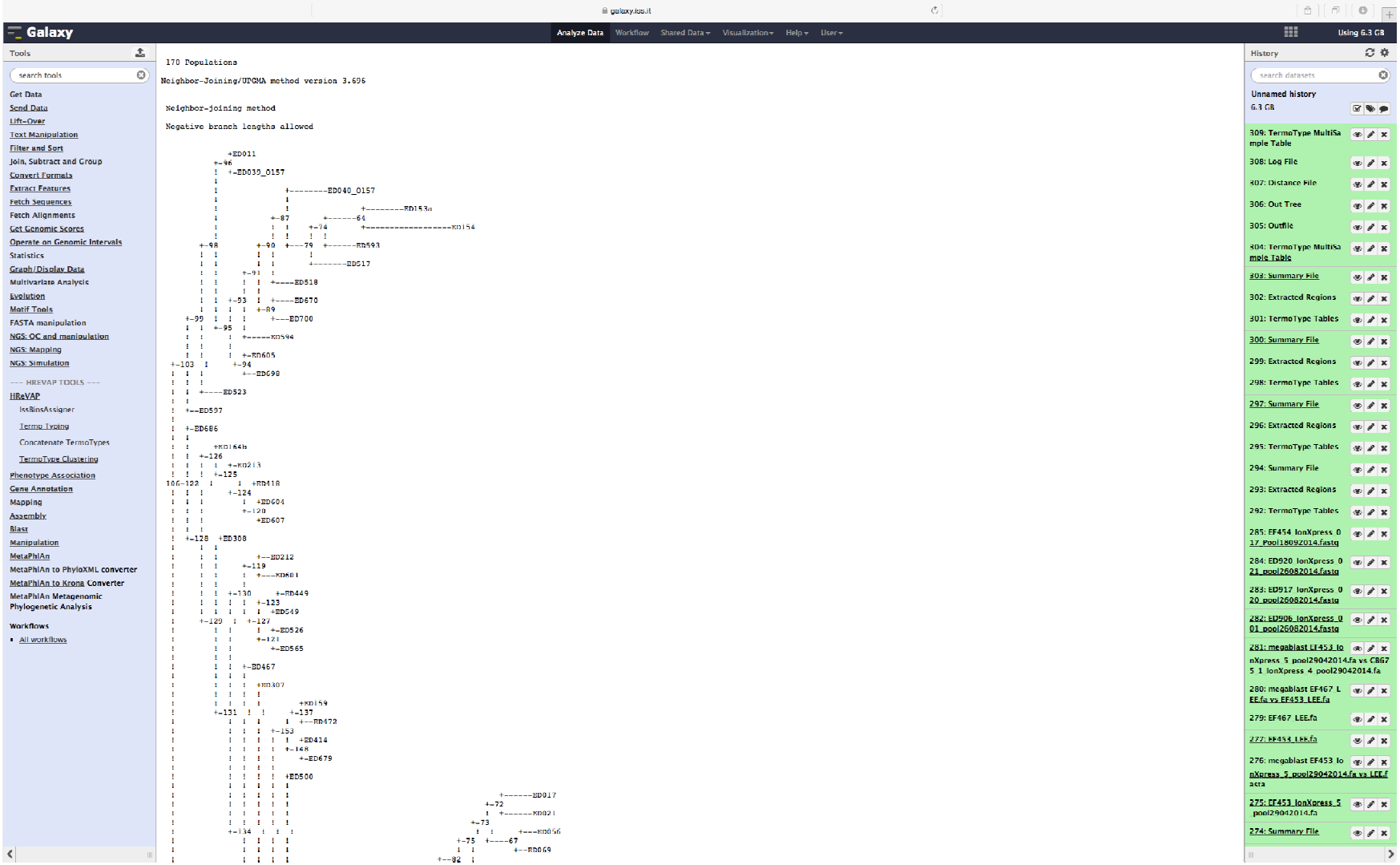
FREE, USER-FRIENDLY WEB INTERFACE

Galaxy

- *de novo* assembly tools
- BLAST search of interesting genes
- Alignment of sequences, production of VCF files, production of dendrograms

OPEN SOURCE, USER-FRIENDLY WEB INTERFACE, OPEN FOR INTRODUCTION OF CUSTUMIZED TOOLS, ELECTION PLATFORM FOR DEVELOPING AND SHARING OF NEW TOOLS

Galaxy instance @ ISS



Example of data analysis: O26 cluster of cases in 2013

22 Italian strains

In house sequencing

NCBI

9 strains

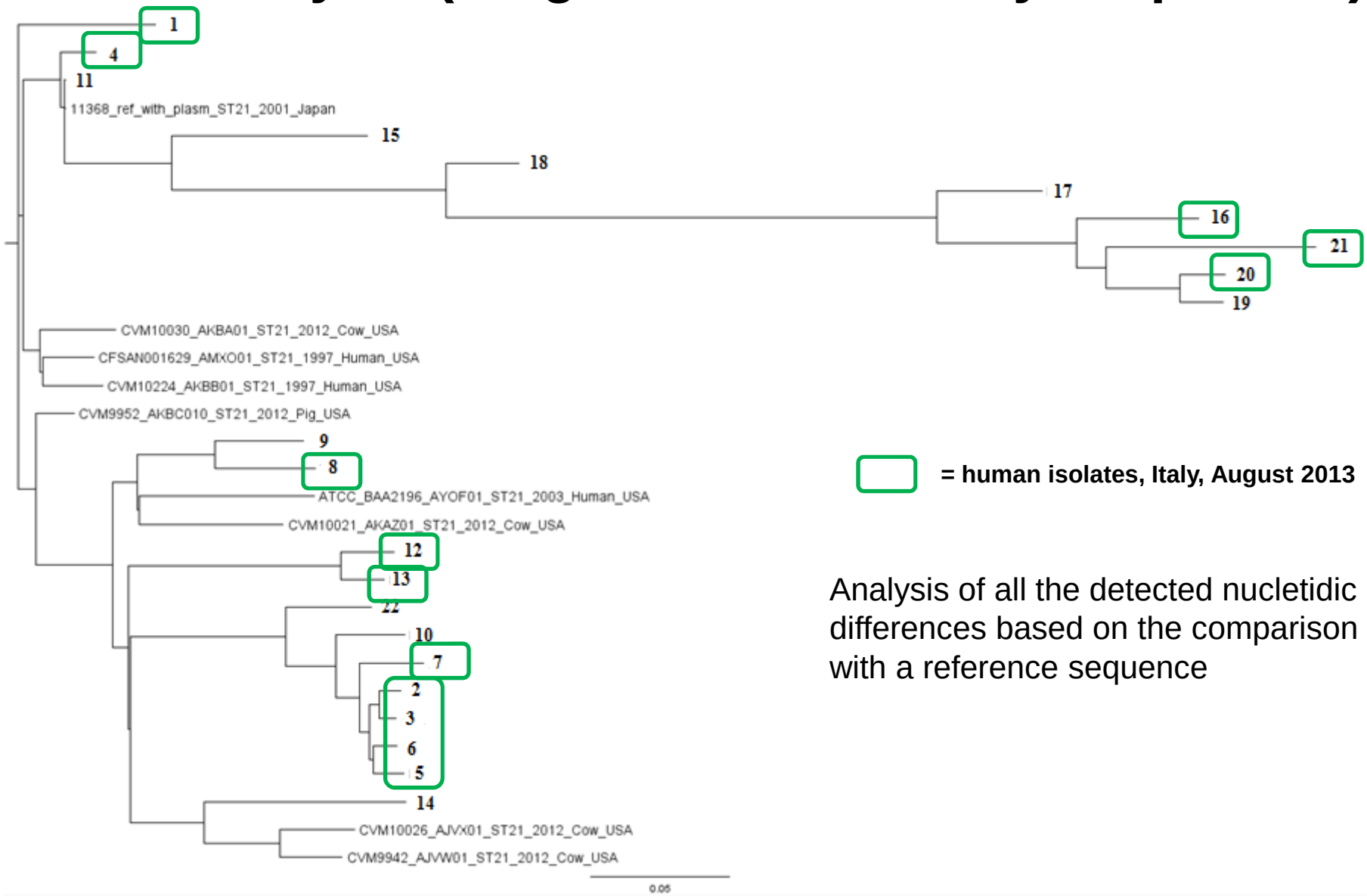
GenBank data

- Virulence genes
- Multi-Locus Sequence Typing (MLST)
- Single Nucleotidic Polymorphisms (SNPs) based phylogenetic tree

	ST21	ST29	*vtx1*	*vtx2*	*vtx-*
13 Human isolates from Italy August 2013	8	5	-	12	1

At least two different strains involved in the cases occurred in Italy in 2013

SNPs analysis (Single Nucleotidic Polymorphisms)



= human 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

Center for Genomic Epidemiology

NDtree

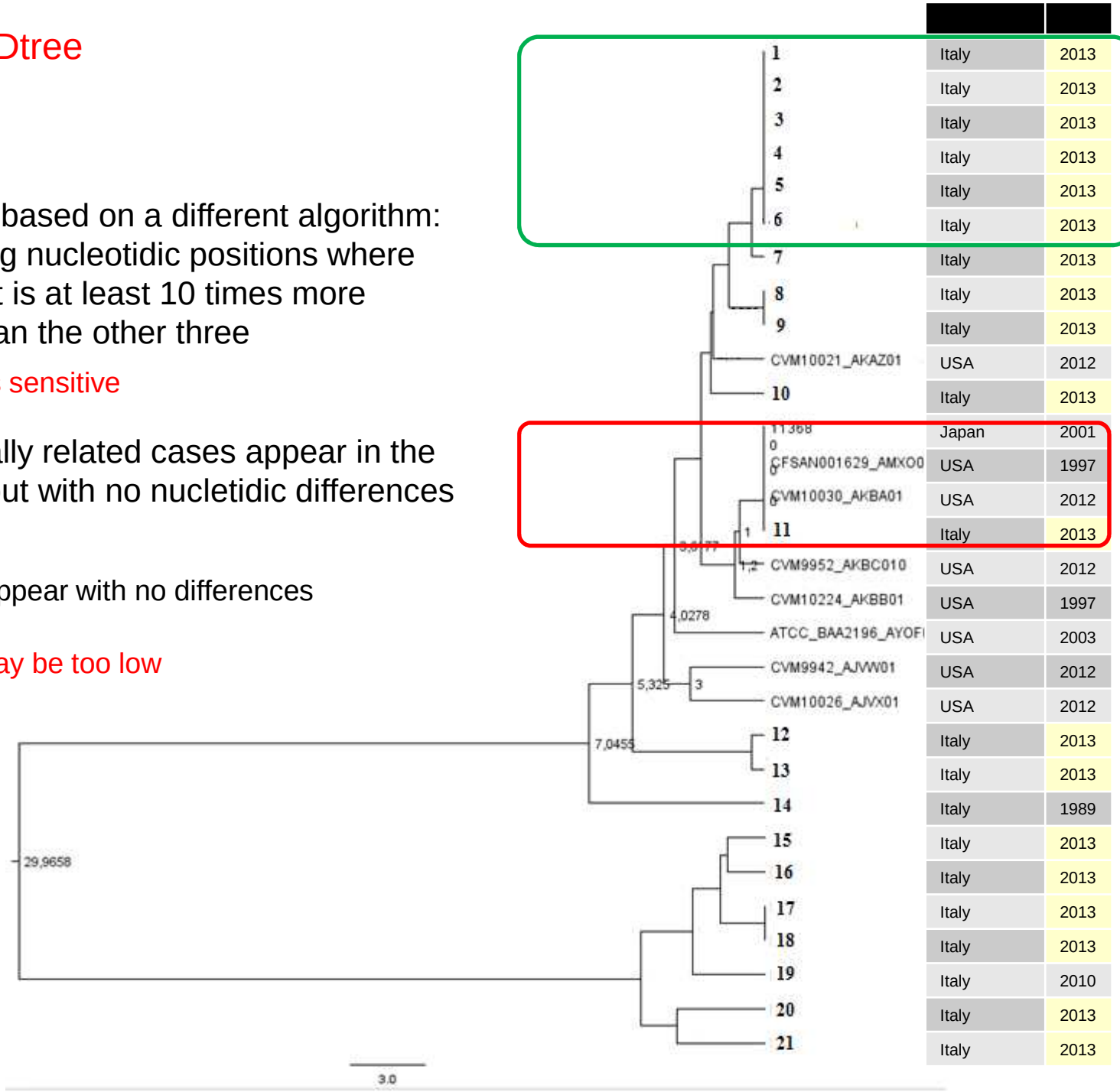
SNPs analysis based on a different algorithm: only considering nucleotidic positions where the assigned nt is at least 10 times more represented than the other three

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



1	Italy	2013
2	Italy	2013
3	Italy	2013
4	Italy	2013
5	Italy	2013
6	Italy	2013
7	Italy	2013
8	Italy	2013
9	Italy	2013
CVM10021_AKAZ01	USA	2012
10	Italy	2013
11368	Japan	2001
0 CFSAN001629_AMXO0	USA	1997
0 CVM10030_AKBA01	USA	2012
11	Italy	2013
CVM9952_AKBC010	USA	2012
CVM10224_AKBB01	USA	1997
ATCC_BAA2196_AYOF	USA	2003
CVM9942_AJVW01	USA	2012
CVM10026_AJVX01	USA	2012
12	Italy	2013
13	Italy	2013
14	Italy	1989
15	Italy	2013
16	Italy	2013
17	Italy	2013
18	Italy	2013
19	Italy	2010
20	Italy	2013
21	Italy	2013

Issues to be addressed

- **Data production still needs to be streamlined**

  Reference laboratories only actively produce data as of today (6 NRLs in our network)

- **Cross-platform compatibility**

  Different platforms = different errors rates and types

- **Intrinsic quality of the sequence reads at the nucleotidic level**

  Filtering algorythms to be developed and harmonized

- **Refinement of existing tools for data analysis and development of new ones**

  Need for new approaches to typing

- **Need for education in bioinformatics**

- **Computationally intense data analysis**

  Accessibility of bioinformatic tools via open-source servers

- **Massive data storage and transfer**

  What data should be stored? Cloud storage?

Problem solving