Report of the first inter-laboratory exercise on Whole Genome Sequencing of Shiga toxin-producing *Escherichia coli* strains 2017-2018 (PT-WGS1)

Edited by:

Silvia Arancia, Susan Babsa, Gianfranco Brambilla, Paola Chiani, Clarissa Ferreri, Fabio Galati, Arnold Knijn, Antonella Maugliani, Valeria Michelacci, Fabio Minelli, Stefano Morabito, Rosangela Tozzoli

Acknowledgments: *Thanks to Joao Andrè Carrico for the support on the INNUENDO tools and for the visualization of the Minimum Spanning Tree on Phyloviz 2.0*

1. INTRODUCTION

The duties of the EU Reference Laboratory for *E. coli* (EURL-VTEC) include the development of analytical methods and the coordination of their application by the National Reference Laboratories (NRLs) for *E. coli* in the EU and EFTA Member States, EU Candidate Countries and certain third countries. One of the key actions in the capacity building towards the use of analytical methodologies is the organization of inter-laboratories studies.

The rapid development of next generation sequencing platforms and the parallel development of bioinformatics tools for NGS data management and analyses is making the genome sequence-based investigation a realistic alternative to conventional molecular typing of bacterial isolates. Many laboratories in the EU have started shifting their analytical procedures from conventional methods to whole genome sequencing (WGS), which has become a valuable alternative to Pulsed Field Gel Electrophoresis (PFGE) for molecular surveillance of *E. coli* infections. Nevertheless, no standard procedures are available, yet, for the application of such technique for *E. coli* characterization and typing.

In November 2017 the EURL-VTEC organized for the first time a voluntary inter-laboratory exercise on WGS of pathogenic *E. coli*, to be run in parallel to the sixth study organized by EURL-VTEC on typing pathogenic *E. coli* through PFGE for the benefit of the network of NRLs for *E. coli* (PT-PFGE6). This inter-laboratory exercise was extended also to Italian Official Laboratories (OLs) and this document is meant to present the results of this study.

2. DESIGN AND OBJECTIVES OF THE STUDY

The study consisted in the production of the whole genome sequences of six STEC strains shipped as soft agar cultures, by using the preferred DNA extraction protocol and Next Generation Sequencing technology and procedure. Following the evaluative purposes of this exercise, each participant was requested to apply the protocols in use for the routine workflows in the respective laboratory.

The **objectives** of the study were:

- to evaluate the quality parameters of the sequences produced and their effect on the WGS-based characterisation of STEC
- to evaluate the inter laboratory and platform variability in terms of SNPs in the genomes produced.

3. PARTICIPANTS

A total of 21 Laboratories including 18 NRLs and 3 Italian OLs voluntarily participated in the study. Each participant received its own individual Laboratory code, reported in the result tables and figures.

The NRLs participating in the study were:

- Austria, *Institut für Medizinische Mikrobiologie und Hygiene*, AGES
- Belgium, Scientific Institute of Public Health, *Direction Opérationnelle Maladies Transmissibles et Infectieuses*
- Denmark, FVST, *Mikrobiologisk Laboratorium*
- Finland, Finnish Food Safety Authority Evira, Research and Laboratory Service Department, Microbiology Research Unit, Helsinki
- Finland, Finnish Food Safety Authority Evira Veterinary Bacteriology Research Unit, Kuopio
- Ireland, HSE Community Healthcare East, Public Health Laboratory, Cherry Orchard Hospital, Ballyfermot, Dublin
- Ireland, Veterinary Public Health Regulatory Laboratory, Department of Agriculture, Food and the Marine
- Italy, *Istituto Superiore di Sanità*
- Latvia, Microbiological Division, Laboratory of Food and Environmental Investigations, Institute of Food Safety, Animal Health and Environment (BIOR)
- Luxembourg, *Ministère de l'Agriculture, de la Viticulture et de la Protection des consommateurs, Administration des services vétérinaires*, LMVE
- Poland, National Veterinary Research Institute, Department of Hygiene of food of animal origin
- Portugal, *Laboratório de Microbiologia dos Alimentos, Instituto Nacional de Investigação Agrária e Veterinária, I.P, Unidade Estratégica de Investigação e Serviços de Tecnologia e Segurança Alimentar* (LNIV)
- Spain, *Unidad Microbiología - Centro Tecnológico Agroalimentario de Lugo*
- Sweden, *Livsmedelsverket*/The National Food Agency
- Sweden, National Veterinary Institute (SVA), Dept of Bacteriology
- The Netherlands, RIVM, Centre for Zoonoses and Environmental Microbiology
- The Netherlands, Food and Consumer Product Safety Authority (NVWA)
- UK, FW&E Laboratory – London, Public Health England

The Italian OLs participating in the study were:

- IZS *Abruzzo e Molise "G. Caporale", Batteriologia e Igiene delle produzioni lattiero casearie, Laboratorio Nazionale di Riferimento per Campylobacter,* Teramo
- IZS *Sardegna, Laboratorio di Microbiologia e Terreni Colturali,* Sassari
- IZS *Piemonte Liguria e Valle d'Aosta, Laboratorio Controllo Alimenti,* Torino

4. MATERIALS AND METHODS

4.1. Sample preparation

The study was carried out on a set of 6 *E. coli* test strains, corresponding to the same samples sent for PT-PFGE6. The cultures were prepared between October 30th and November 8th 2017 to be used as test material. They consisted in freshly prepared bacterial cultures seeded into soft (0.3 %) nutrient agar in 2 ml glass vials, which were incubated for 18 hours at 37 °C ± 1 °C and labelled with numbers from 1 to 6 followed by the specific lab code (*e.g.* Strain 1 Lxxx). The homogeneity of the test strains was assessed on November 9th 2017 by testing two randomly selected sets of strains for the presence of known genetic characteristics (Table 1). The test samples were stored at room temperature until November 13th, when they were sent to the participating laboratories by courier.

As for the stability of the samples, previous experiences supported the assumption that the time range between the preparation of the specimens and the deadline for submission of results was short enough to assure the stability of the strains.

Table 1: Characteristics of the test strains

Number	Strain	Serotype	MLST	Genotype
1	ED 56	O26:H11	21	*stx1 eae ehxA*
2	ED 258	O26:H11	21	*stx1 eae*
3	ED 477	O26:H11	21	*stx2 eae ehxA*
4	ED 600	O26:H11	21	*stx1 eae*
5	ED 1014	O26:H11	21	*stx2 eae ehxA*
6	ED 1104	O26:H11	21	*stx2 eae ehxA*

4.2. Laboratory methods

The laboratories that agreed to participate in this voluntary exercise were requested to sequence the whole genome of the six test strains using the DNA extraction protocol and Next Generation Sequencing technology and procedures applied to the routine workflow.

4.3. Collection and elaboration of the results

The participants were requested to submit only the ".fastq" files, outputs of the whole genome sequencing, without trimming and with no additional bioinformatics analyses carried out. For each sequence submitted by each laboratory, the files submitted and all the results of the analyses performed were renamed according to the strain number and the lab code (*e.g*. Strain1_Lxxx).

The participants were asked to submit the data through an instance of the suite distributed by the Integrated Rapid Infectious Disease Analysis (IRIDA) project (Matthews *et al.*, 2018), installed on a webserver at ISS (https://irida.iss.it). The choice to submit the results as unzipped ".fastq" files or as compressed files in the format ".fastq.gz" was left open to the participants.

Specific username and password for accessing the IRIDA-ISS instance were provided to each participant by email.

4.4. Analysis of the results

The sequence files were collected and analysed using the tools present in the ARIES webserver developed by the EURL-VTEC (https://www.iss.it/site/aries).

All the submitted files were processed through an automatic pipeline developed by the EURL-VTEC (EURL VTEC WGS PT, Galaxy version 1.0) performing the following operations: quality check, trimming, assembly, assembly statistics, Multi Locus Sequence Typing (MLST) with the conventional scheme of seven housekeeping genes (Wirth *et al.*, 2006), serotyping and virulotyping. Details on the tools and parameters used in the pipeline

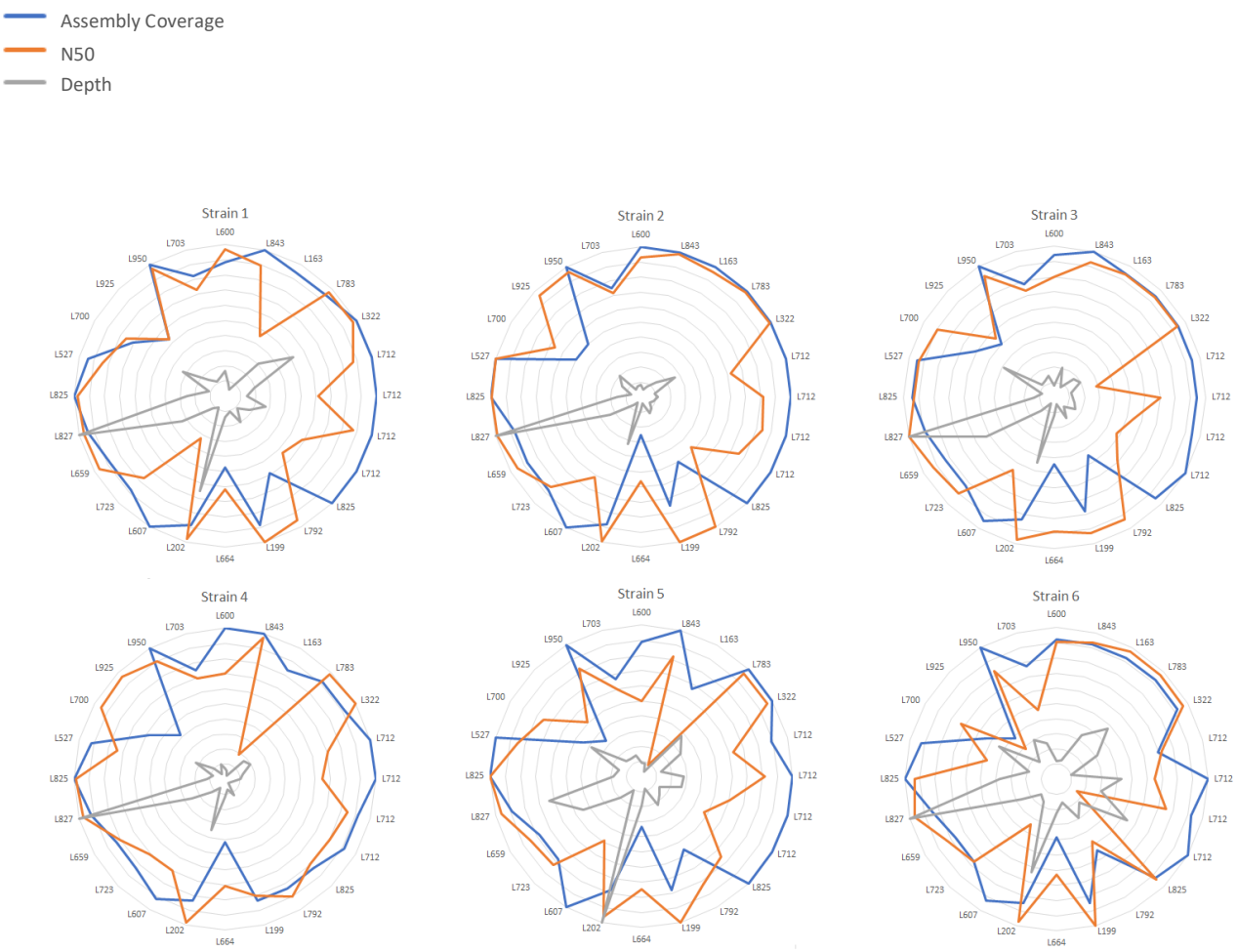
are illustrated in Annex 1. The assembled contigs were then used to detect differences at the whole genome level through pipelines developed for phylogenetic analysis. The methodology applied was the same for all the data analysed. The aim of this step was to estimate the inter-laboratory variability in the whole genome sequence of the same strain.
The tools used reflect those available online as open source software and published on peer reviewed journals. It is possible that errors have been identified during this study, which may not have occurred by using different analytical tools.

5. RESULTS

5.1. Overview of the data panel

All the participating laboratories uploaded the raw .fastq files on the IRIDA-ISS webserver.
One laboratory (L712) returned the results of four different sequencing runs per test strain, each run being coded with a different suffix. The results of each of these runs are reported separately for this lab, keeping the rune labels provided by the uploader (L712_Qa, L712_Qb, L712_R, L712_R2).
Among the 21 participants, 17 laboratories provided data as paired-end reads and the remaining four as single-end reads.
The results of the depth and assembly statistics including the N50 and the total length of assembled contigs compared to the expected length of 5 Mb (assembly coverage) were calculated and used to build the graphs presented in Figure 1. This analysis allowed to have a general overview at a glance of the whole data panel, especially in terms of depth of sequencing and of N50 achieved after assembly of the reads. In detail, strain-to-strain variation could be observed, highlighting different levels of sequencing achieved by the same lab when sequencing the various test strains. Moreover, differences could be identified also among the test strains, with strains 5 and 6 presenting the most heterogeneous patterns of results achieved with the sequences produced by the participating labs. The high diversity observed in this analysis reflects the differences obtained in all the downstream analysis, presented in the next paragraphs.

Figure 1. Analysis of the sequencing depth, assembly N50 and assembly coverage achieved with the results produced by the participating labs.



5.2. Sequencing depth

The sequencing depth achieved for the genome of each test isolate by the participating laboratories was estimated on the mean coverage observed for the seven genes part of the conventional scheme for Multi Locus Sequence Typing (Wirth *et al*., 2006). The results are presented in Table 1. The range of depth calculated across all the labs and all the genomes was comprised between 6.6x and 309.5x, with a mean value of 43.6x and a median value of 30.7x. Seven laboratories achieved a depth of 30x or higher for all the six strains tested. It is interesting to note that differences in depth values could be observed also in sequences produced by the same laboratory.

Table 1. Sequencing depth calculated for the sequences submitted by the participating labs.

The following legend applies to the boxes' colours:

Green: depth ≥30x

Yellow: 20≤depth<30

Orange: 10≤depth<20

Red: depth<10

Lab code	Strain1	Strain2	Strain3	Strain4	Strain5	Strain6
L163	8,5	25,5	18,9	7,6	6,6	37,5
L199	17,7	13,0	10,6	21,1	14,9	17,7
L202	109,3	100,3	106,3	105,6	181,3	70,9
L322	87,0	80,8	46,8	59,0	53,9	35,0
L527	19,3	20,7	20,6	25,2	28,1	20,9
L600	28,1	25,7	18,8	22,2	16,3	13,5
L607	14,4	13,1	11,2	20,0	18,7	19,3
L659	55,6	72,8	123,8	76,7	80,6	29,2
L664	23,6	26,9	27,9	34,8	34,7	24,9
L700	54,1	44,2	92,3	68,3	69,2	48,8
L703	20,9	22,7	35,1	31,2	25,6	27,7
L712_Qa	33,1	29,7	26,5	38,1	24,0	11,5
L712_Qb	23,6	35,6	28,3	31,9	50,5	48,0
L712_R	46,4	26,9	31,8	18,0	49,1	34,3
L712_Rb	30,7	20,6	37,0	13,4	25,1	60,2
L723	19,4	21,7	29,9	35,5	35,3	15,5
L783	51,2	43,9	41,7	53,0	68,2	53,0
L792	33,5	37,6	37,6	37,1	39,4	33,0
L825	43,0	48,9	31,9	35,3	33,5	41,9
L825b	18,6	31,3	21,3	10,3	30,6	23,9
L827	167,6	309,5	237,2	305,0	114,1	111,6
L843	13,6	17,0	48,8	17,4	16,1	14,5
L925	25,9	62,3	28,3	31,5	26,0	17,7
L950	18,9	18,9	29,9	12,1	24,9	33,2

5.3. Multi Locus Sequence Typing – seven loci scheme

The Sequence Type (ST) could be correctly identified in all the strains from the sequences produced by 10 out of the 18 participating laboratories. As a whole, out of the 144 submitted and analysed sequences, for 15 it was not possible to determine the ST (STNF, Sequence Type Not Found), as the combination of alleles found did not correspond to any known ST, and from eleven the identification of ST21 was uncertain.

Only three laboratories submitted more than one genome for which it was not possible to predict the correct ST (L163, L199 and L659).

Table 2. Multi Locus Sequence Types identified in the sequences submitted by the participating labs.

The following legend applies to the boxes' colours:

Green: exact match of all the alleles, exact ST prediction

Orange: uncertain prediction of the correct ST

Red: Sequence Type Not Found (STNF), due to an unknown combination of the alleles of the seven loci

* next to the ST indicates that there were mismatches against at least one of the alleles.

? indicates that there was uncertainty in at least one of the alleles.

Lab code	Strain1	Strain2	Strain3	Strain4	Strain5	Strain6
True value	ST21	ST21	ST21	ST21	ST21	ST21
L163	STNF?	ST21	ST21	STNF*?	STNF*?	ST21
L199	STNF	ST21*?	ST21*?	ST21?	STNF	STNF
L202	ST21	ST21	STNF	ST21	ST21	ST21
L322	ST21	ST21	ST21	ST21	ST21	ST21
L527	ST21	ST21	ST21	ST21	ST21	ST21
L600	ST21	ST21	ST21?	ST21	ST21	ST21?
L607	ST21?	STNF*	ST21	ST21	ST21	ST21
L659	STNF	ST21	STNF	STNF	ST21	STNF
L664	ST21	ST21	ST21?	ST21	ST21	ST21
L700	ST21	ST21	ST21	ST21	ST21	ST21
L703	ST21	ST21	ST21	ST21	ST21	STNF*?
L712_Qa	ST21	ST21	ST21	ST21	ST21	ST21
L712_Qb	ST21	ST21	ST21	ST21	STNF*?	ST21
L712_R	ST21	ST21	ST21	ST21	ST21	ST21
L712_Rb	ST21	ST21	ST21	ST21?	ST21	ST21
L723	ST21	ST21	ST21	STNF*?	ST21	ST21?
L783	ST21	ST21	ST21	ST21	ST21	ST21
L792	ST21	ST21	ST21	ST21	ST21	ST21
L825	ST21	ST21	ST21	ST21	ST21	ST21
L825b	ST21	ST21	ST21	ST21?	ST21	ST21
L827	ST21	ST21	ST21	ST21	ST21	ST21
L843	ST21	ST21	ST21	ST21?	ST21	ST21
L925	ST21	ST21	ST21	ST21	ST21	ST21
L950	ST21	ST21	ST21	ST21	ST21	ST21

5.4. Serotyping

Serotyping of the WGS submitted by the participating labs was assayed using the EURL-VTEC WGS PT pipeline described in Annex 1. The H-type could be correctly identified in all the sequences. The serogroup could not be identified only from six sequences, one per test strain. In detail, two of the sequencing runs submitted by L712 presented problems in serogrouping of five of the six sequences (Table 3).

Table 3. *In silico* serotyping of the sequences submitted by the participating labs

The following legend applies to the boxes' colours :

Green: exact serotype prediction

Red: Errors detected in serotype prediction.

Lab code	Strain1	Strain2	Strain3	Strain4	Strain5	Strain6
True value	O26:H11	O26:H11	O26:H11	O26:H11	O26:H11	O26:H11
L163	O26:H11	O26:H11	O26:H11	O26:H11	O?:H11	O26:H11
L199	O26:H11	O26:H11	O26:H11	O26:H11	O26:H11	O26:H11
L202	O26:H11	O26:H11	O26:H11	O26:H11	O26:H11	O26:H11
L322	O26:H11	O26:H11	O26:H11	O26:H11	O26:H11	O26:H11
L527	O26:H11	O26:H11	O26:H11	O26:H11	O26:H11	O26:H11
L600	O26:H11	O26:H11	O26:H11	O26:H11	O26:H11	O26:H11
L607	O26:H11	O26:H11	O26:H11	O26:H11	O26:H11	O26:H11
L659	O26:H11	O26:H11	O26:H11	O26:H11	O26:H11	O26:H11
L664	O26:H11	O26:H11	O26:H11	O26:H11	O26:H11	O26:H11
L700	O26:H11	O26:H11	O26:H11	O26:H11	O26:H11	O26:H11
L703	O26:H11	O26:H11	O26:H11	O26:H11	O26:H11	O26:H11
L712_Qa	O?:H11	O26:H11	O26:H11	O?:H11	O26:H11	O?:H11
L712_Qb	O26:H11	O?:H11	O?:H11	O26:H11	O26:H11	O26:H11
L712_R	O26:H11	O26:H11	O26:H11	O26:H11	O26:H11	O26:H11
L712_Rb	O26:H11	O26:H11	O26:H11	O26:H11	O26:H11	O26:H11
L723	O26:H11	O26:H11	O26:H11	O26:H11	O26:H11	O26:H11
L783	O26:H11	O26:H11	O26:H11	O26:H11	O26:H11	O26:H11
L792	O26:H11	O26:H11	O26:H11	O26:H11	O26:H11	O26:H11
L825	O26:H11	O26:H11	O26:H11	O26:H11	O26:H11	O26:H11
L825b	O26:H11	O26:H11	O26:H11	O26:H11	O26:H11	O26:H11
L827	O26:H11	O26:H11	O26:H11	O26:H11	O26:H11	O26:H11
L843	O26:H11	O26:H11	O26:H11	O26:H11	O26:H11	O26:H11
L925	O26:H11	O26:H11	O26:H11	O26:H11	O26:H11	O26:H11
L950	O26:H11	O26:H11	O26:H11	O26:H11	O26:H11	O26:H11

5.5. Virulotyping

The identification of the presence of the main virulence genes of STEC (*stx1*, *stx2*, *eae* and *ehxA*) in the genome of the test strains was assayed by using the EURL-VTEC WGS PT pipeline described in Annex 1. The results are reported in Table 4.

From the analysis of the results, the inversion of the strains 3 and 4 submitted by L792, and the inversion of strains 4 and 6 submitted by L712_Qa was suspected (Table 4).

In addition, lack of identification of *stx2* in *stx2*-positive samples was observed in four sequences (strain 5 for L163 and L925 and strain 6 for L703 and L792).

One participant, L600, submitted two sequences for strain 1, following the observation of two slightly different morphologies on agar plates. The additional sequence provided was

tested in all the pipelines and did not display any differences in the results, apart from the absence of *ehxA*. An asterisk (*) symbol has been used to flag this result in Table 4.

Table 4. *In silico* virulotyping of the sequences submitted by the participating labs

The following legend applies to the boxes' colours:

Green: exact prediction

Red: Error detected in the prediction

*: absence of identification of *ehxA* gene in one of the two sequences provided for the same strain.

	Strain 1				Strain 2			
Lab code	***stx1***	***stx2***	***eae***	***ehxA***	***stx1***	***stx2***	***eae***	***ehxA***
True value	**+**	**-**	**+**	**+**	**+**	**-**	**+**	**-**
L163								
L199								
L202								
L322								
L527								
L600				*				
L607								
L659								
L664								
L700								
L703								
L712_Qa								
L712_Qb								
L712_R								
L712_Rb								
L723								
L783								
L792								
L825								
L825b								
L827								
L843								
L925								
L950								

	Strain 3				Strain 4			
Lab code	***stx1***	***stx2***	***eae***	***ehxA***	***stx1***	***stx2***	***eae***	***ehxA***
True value	**-**	**+**	**+**	**+**	**+**	**-**	**+**	**-**
L163								
L199								
L202								
L322								
L527								
L600								
L607								
L659								
L664								
L700								
L703								
L712_Qa								
L712_Qb								
L712_R								
L712_Rb								
L723								
L783								
L792								
L825								
L825b								
L827								
L843								
L925								
L950								

	Strain 5				Strain 6			
Lab code	***stx1***	***stx2***	***eae***	***ehxA***	***stx1***	***stx2***	***eae***	***ehxA***
True value	**-**	**+**	**+**	**+**	**-**	**+**	**+**	**+**
L163								
L199								
L202								
L322								
L527								
L600								
L607								
L659								
L664								
L700								
L703								
L712_Qa								
L712_Qb								
L712_R								
L712_Rb								
L723								
L783								
L792								
L825								
L825b								
L827								
L843								
L925								
L950								

5.6. Phylogenetic analysis

The contigs assembled have been compared all-against-all through the use of different pipelines to investigate the variability of the sequences submitted.

The ultimate purpose of this exercise was to investigate the differences arising when different laboratories sequence the same strain, mimicking the possibility that an outbreak strain is sequenced at different labs. Such inter-laboratory variability was evaluated by applying the following methodologies:

- SNPs comparison through a reference-free, kmer-based approach (whole genome SNPs and core SNPs)
- whole genome MLST (wgMLST). INNUENDO Scheme 7601 genes
- core genome MLST (cgMLST), INNUENDO Scheme 2360 core genes
- dynamic core genome MLST (dynamic cgMLST), selecting core genes shared by all the sequences analyzed.

Details of these methodologies are described in Annex 1.

Each analysis was carried out either before or after assembly optimization (Annex 1). Only the assemblies satisfying the preliminary quality check (Annex 1) were used for assembly optimization and for the corresponding phylogenetic analyses. Moreover, the pipeline used for assembly optimization only accepted paired-end reads in input and for this reason all single-end reads were not optimized or used for the following analysis, regardless their quality. The dendrograms obtained through SNPs analysis were visualized and coloured by using FigTree software v 1.4.3 (http://tree.bio.ed.ac.uk/software/figtree/), while the Minimum Spanning Trees (MST) obtained through wgMLST, cgMLST and dynamic cgMLST were visualised and coloured by using Phyloviz online 2.0 (Nascimento *et al.*, 2017).

5.6.1. Whole genome SNPs analysis through a kmer-based reference-free approach

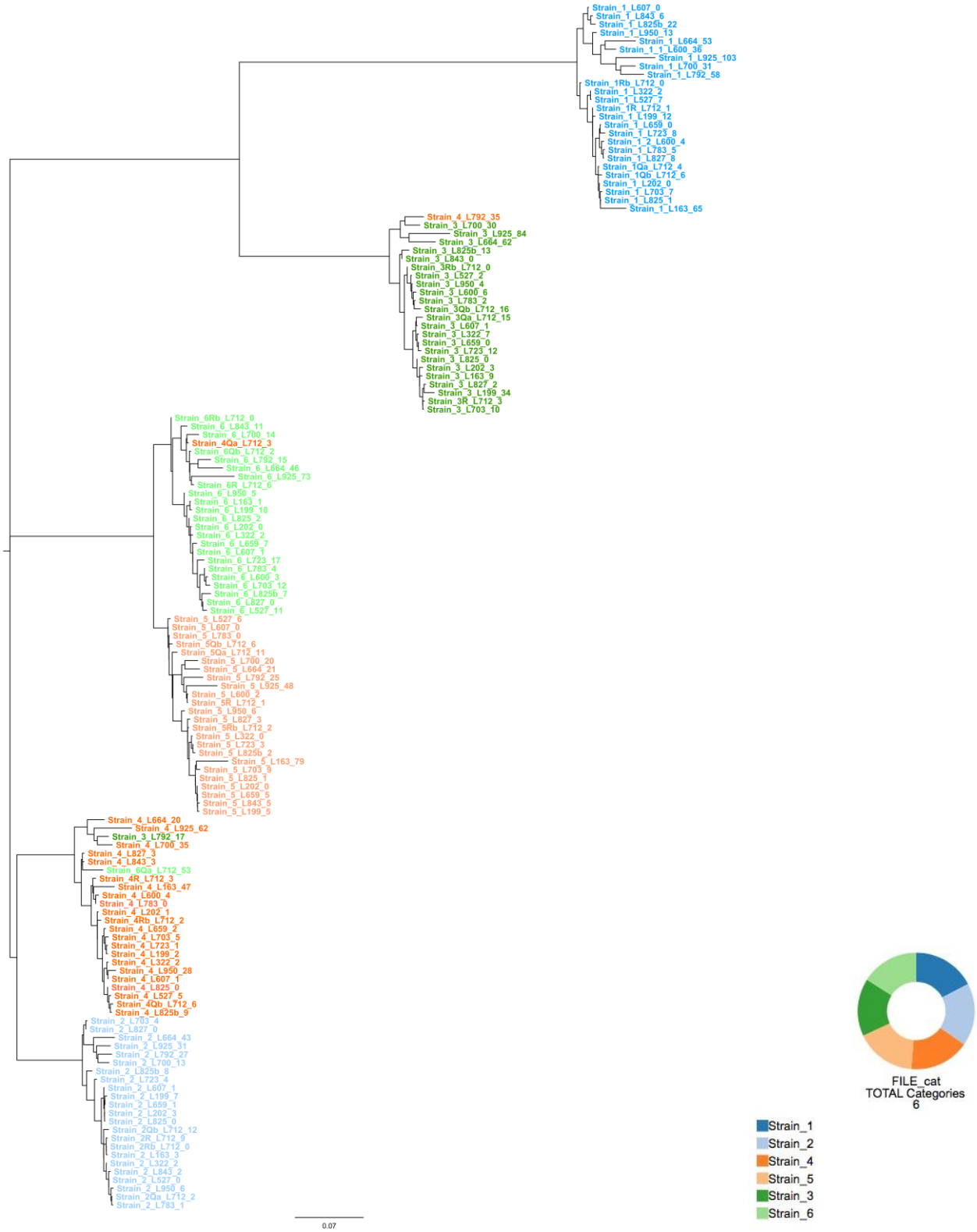
The contigs assembled from all the received sequences were used to build a dendrogram based on the Maximum Parsimony clustering algorithm using the ksnp3 pipeline (Gardner *et al.*, 2015). The dendrograms in Figures 2A and 2B correspond to the results of this analysis applied to all the crude assemblies or to optimized assemblies, respectively. The labels include the strain identifier, the lab-code and the strain-specific alleles numbers, each separated by underscores. In detail, the strain-specific alleles numbers reflect the number of different kmer alleles with respect to the previous node.

Six different clusters appeared in both the dendrograms obtained, each containing all the assemblies corresponding to one of the six test strains. Only four exceptions emerged, with assemblies appearing in the clusters corresponding to other strains. This confirmed the hypothesized inversion of strains: strains 3 and 4 for L792 and strains 4 and 6 for the set of sequences labelled “Qa” submitted by L712.

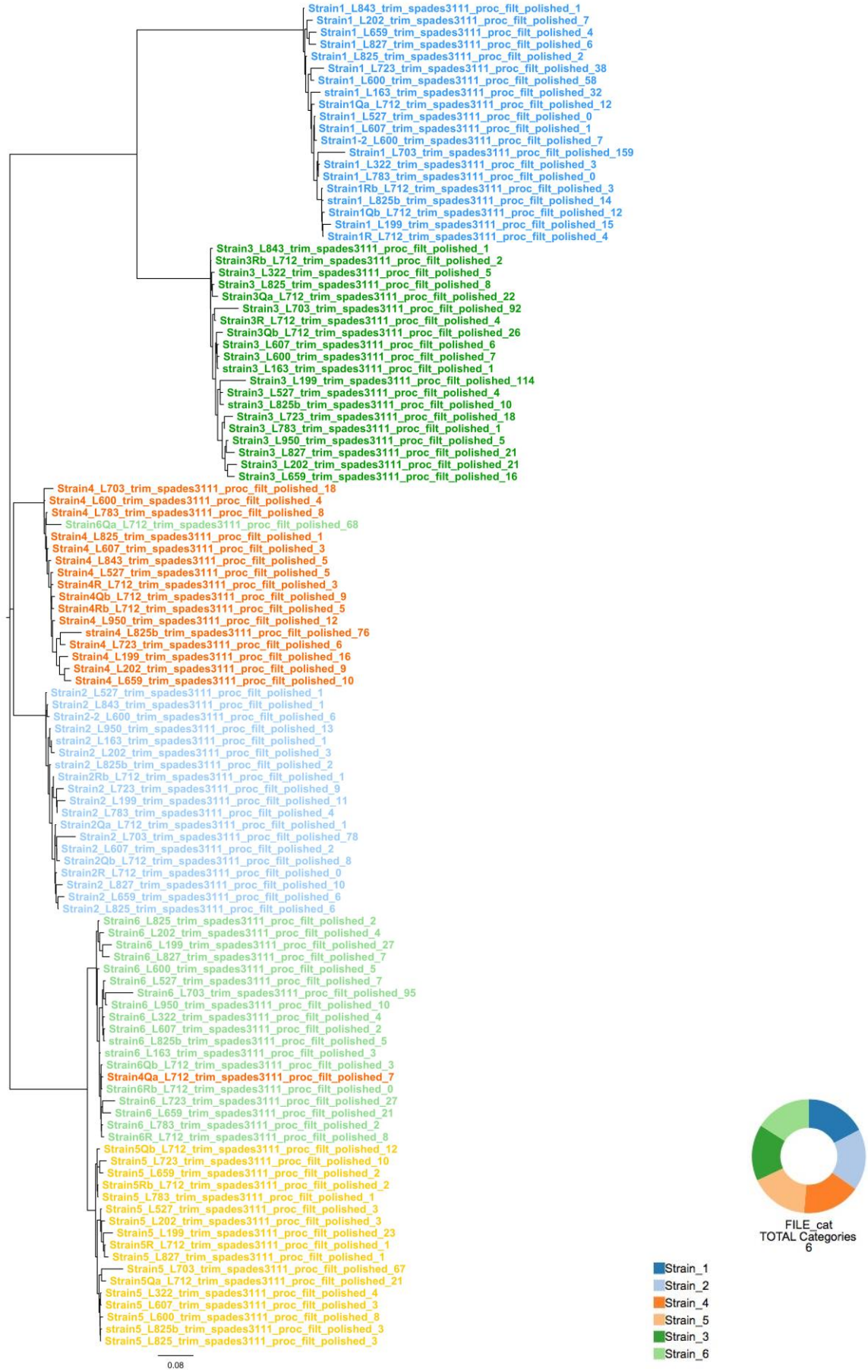
When the analysis was performed on optimized assemblies, the dendrogram topology was unchanged but the intra-cluster variability was lower when compared to the results obtained with crude assembled contigs (Figure 2).

Figure 2. Maximum Parsimony clustering of the ksnp3 analysis of the whole set of crude assembled contigs (2A) and optimized assemblies (2B).

2A



2B


Strain1_L843_trim_spades3111_proc_filt_polished_1
Strain1_L202_trim_spades3111_proc_filt_polished_7
Strain1_L659_trim_spades3111_proc_filt_polished_4
Strain1_L827_trim_spades3111_proc_filt_polished_6
Strain1_L825_trim_spades3111_proc_filt_polished_2
Strain1_L723_trim_spades3111_proc_filt_polished_38
Strain1_L600_trim_spades3111_proc_filt_polished_58
strain1_L163_trim_spades3111_proc_filt_polished_32
Strain1Qa_L712_trim_spades3111_proc_filt_polished_12
Strain1_L527_trim_spades3111_proc_filt_polished_0
Strain1_L607_trim_spades3111_proc_filt_polished_1
Strain1-2_L600_trim_spades3111_proc_filt_polished_7
Strain1_L703_trim_spades3111_proc_filt_polished_159
Strain1_L322_trim_spades3111_proc_filt_polished_3
Strain1_L783_trim_spades3111_proc_filt_polished_0
Strain1Rb_L712_trim_spades3111_proc_filt_polished_3
strain1_L825b_trim_spades3111_proc_filt_polished_14
Strain1Qb_L712_trim_spades3111_proc_filt_polished_12
Strain1_L199_trim_spades3111_proc_filt_polished_15
Strain1R_L712_trim_spades3111_proc_filt_polished_4
Strain3_L843_trim_spades3111_proc_filt_polished_1
Strain3Rb_L712_trim_spades3111_proc_filt_polished_2
Strain3_L322_trim_spades3111_proc_filt_polished_5
Strain3_L825_trim_spades3111_proc_filt_polished_8
Strain3Qa_L712_trim_spades3111_proc_filt_polished_22
Strain3_L703_trim_spades3111_proc_filt_polished_92
Strain3R_L712_trim_spades3111_proc_filt_polished_4
Strain3Qb_L712_trim_spades3111_proc_filt_polished_26
Strain3_L607_trim_spades3111_proc_filt_polished_6
Strain3_L600_trim_spades3111_proc_filt_polished_7
strain3_L163_trim_spades3111_proc_filt_polished_1
Strain3_L199_trim_spades3111_proc_filt_polished_114
Strain3_L527_trim_spades3111_proc_filt_polished_4
strain3_L825b_trim_spades3111_proc_filt_polished_10
Strain3_L723_trim_spades3111_proc_filt_polished_18
Strain3_L783_trim_spades3111_proc_filt_polished_1
Strain3_L950_trim_spades3111_proc_filt_polished_5
Strain3_L827_trim_spades3111_proc_filt_polished_21
Strain3_L202_trim_spades3111_proc_filt_polished_21
Strain3_L659_trim_spades3111_proc_filt_polished_16
Strain4_L703_trim_spades3111_proc_filt_polished_18
Strain4_L600_trim_spades3111_proc_filt_polished_4
Strain4_L783_trim_spades3111_proc_filt_polished_8
Strain6Qa_L712_trim_spades3111_proc_filt_polished_68
Strain4_L825_trim_spades3111_proc_filt_polished_1
Strain4_L607_trim_spades3111_proc_filt_polished_3
Strain4_L843_trim_spades3111_proc_filt_polished_5
Strain4_L527_trim_spades3111_proc_filt_polished_5
Strain4R_L712_trim_spades3111_proc_filt_polished_3
Strain4Qb_L712_trim_spades3111_proc_filt_polished_9
Strain4Rb_L712_trim_spades3111_proc_filt_polished_5
Strain4_L950_trim_spades3111_proc_filt_polished_12
strain4_L825b_trim_spades3111_proc_filt_polished_76
Strain4_L723_trim_spades3111_proc_filt_polished_6
Strain4_L199_trim_spades3111_proc_filt_polished_16
Strain4_L202_trim_spades3111_proc_filt_polished_9
Strain4_L659_trim_spades3111_proc_filt_polished_10
Strain2_L527_trim_spades3111_proc_filt_polished_1
Strain2_L843_trim_spades3111_proc_filt_polished_1
Strain2-2_L600_trim_spades3111_proc_filt_polished_6
Strain2_L950_trim_spades3111_proc_filt_polished_13
strain2_L163_trim_spades3111_proc_filt_polished_1
Strain2_L202_trim_spades3111_proc_filt_polished_3
strain2_L825b_trim_spades3111_proc_filt_polished_2
Strain2Rb_L712_trim_spades3111_proc_filt_polished_1
Strain2_L723_trim_spades3111_proc_filt_polished_9
Strain2_L199_trim_spades3111_proc_filt_polished_11
Strain2_L783_trim_spades3111_proc_filt_polished_4
Strain2Qa_L712_trim_spades3111_proc_filt_polished_1
Strain2_L703_trim_spades3111_proc_filt_polished_78
Strain2_L607_trim_spades3111_proc_filt_polished_2
Strain2Qb_L712_trim_spades3111_proc_filt_polished_8
Strain2R_L712_trim_spades3111_proc_filt_polished_0
Strain2_L827_trim_spades3111_proc_filt_polished_10
Strain2_L659_trim_spades3111_proc_filt_polished_6
Strain2_L825_trim_spades3111_proc_filt_polished_6
Strain6_L825_trim_spades3111_proc_filt_polished_2
Strain6_L202_trim_spades3111_proc_filt_polished_4
Strain6_L199_trim_spades3111_proc_filt_polished_27
Strain6_L827_trim_spades3111_proc_filt_polished_7
Strain6_L600_trim_spades3111_proc_filt_polished_5
Strain6_L527_trim_spades3111_proc_filt_polished_7
Strain6_L703_trim_spades3111_proc_filt_polished_95
Strain6_L950_trim_spades3111_proc_filt_polished_10
Strain6_L322_trim_spades3111_proc_filt_polished_4
Strain6_L607_trim_spades3111_proc_filt_polished_2
strain6_L825b_trim_spades3111_proc_filt_polished_5
strain6_L163_trim_spades3111_proc_filt_polished_3
Strain6Qb_L712_trim_spades3111_proc_filt_polished_3
Strain4Qa_L712_trim_spades3111_proc_filt_polished_7
Strain6Rb_L712_trim_spades3111_proc_filt_polished_0
Strain6_L723_trim_spades3111_proc_filt_polished_27
Strain6_L659_trim_spades3111_proc_filt_polished_21
Strain6_L783_trim_spades3111_proc_filt_polished_2
Strain6R_L712_trim_spades3111_proc_filt_polished_8
Strain5Qb_L712_trim_spades3111_proc_filt_polished_12
Strain5_L723_trim_spades3111_proc_filt_polished_10
Strain5_L659_trim_spades3111_proc_filt_polished_2
Strain5Rb_L712_trim_spades3111_proc_filt_polished_2
Strain5_L783_trim_spades3111_proc_filt_polished_1
Strain5_L527_trim_spades3111_proc_filt_polished_3
Strain5_L202_trim_spades3111_proc_filt_polished_3
Strain5_L199_trim_spades3111_proc_filt_polished_23
Strain5R_L712_trim_spades3111_proc_filt_polished_1
Strain5_L827_trim_spades3111_proc_filt_polished_1
Strain5_L703_trim_spades3111_proc_filt_polished_67
Strain5Qa_L712_trim_spades3111_proc_filt_polished_21
Strain5_L322_trim_spades3111_proc_filt_polished_4
Strain5_L607_trim_spades3111_proc_filt_polished_3
Strain5_L600_trim_spades3111_proc_filt_polished_8
strain5_L825b_trim_spades3111_proc_filt_polished_3
Strain5_L825_trim_spades3111_proc_filt_polished_3
0.08
FILE_cat
TOTAL Categories
6
Strain_1
Strain_2
Strain_4
Strain_5
Strain_3
Strain_6

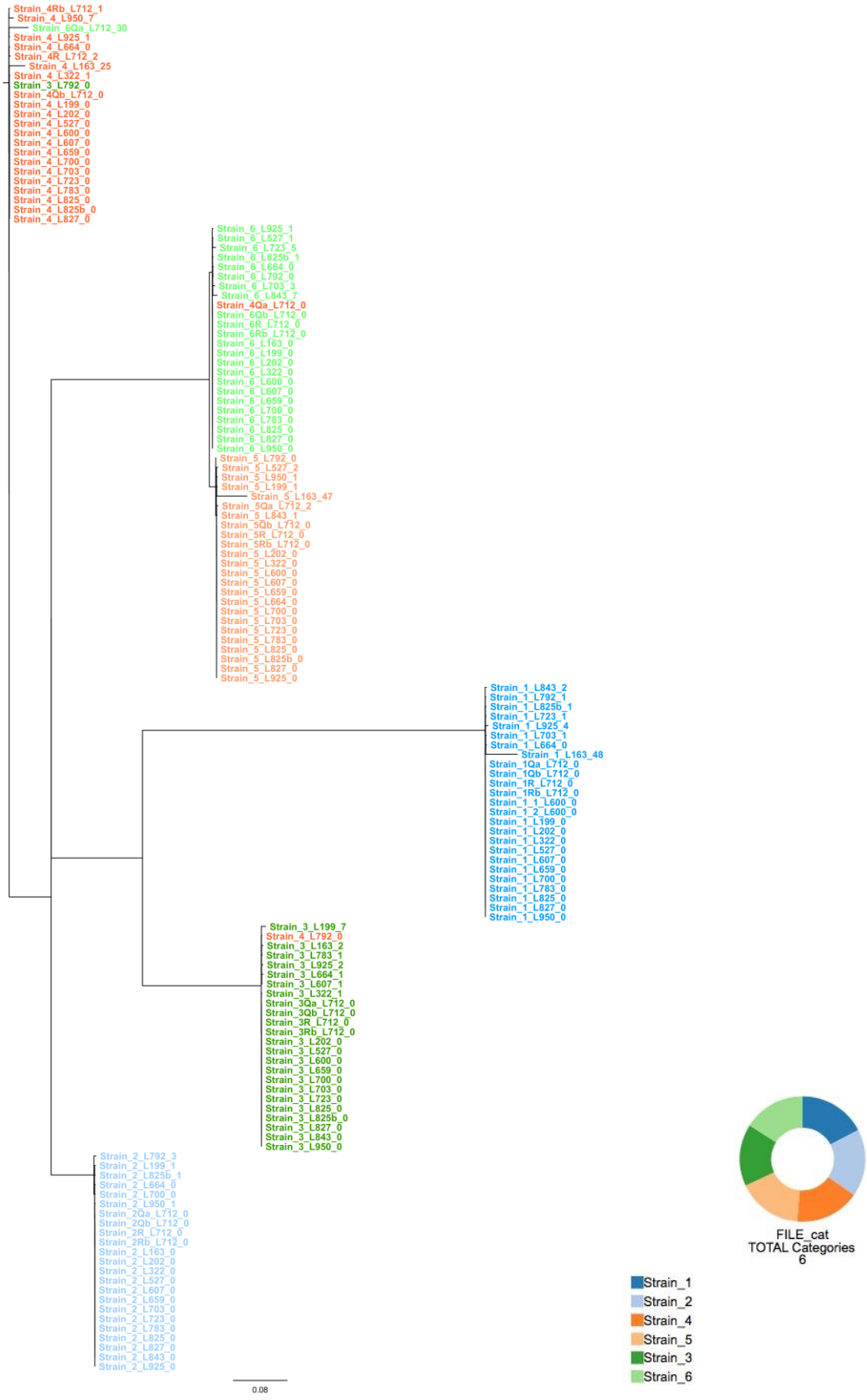
5.6.2 Core genome SNPs analysis (kmer-based reference-free approach)

The ksnp3 analysis was repeated by considering only the core kmers, shared by 90 % of the sequences analysed, as detailed in Annex 1. The results are presented in Figures 3A and 3B, obtained with crude assembled contigs and with optimized assemblies, respectively.

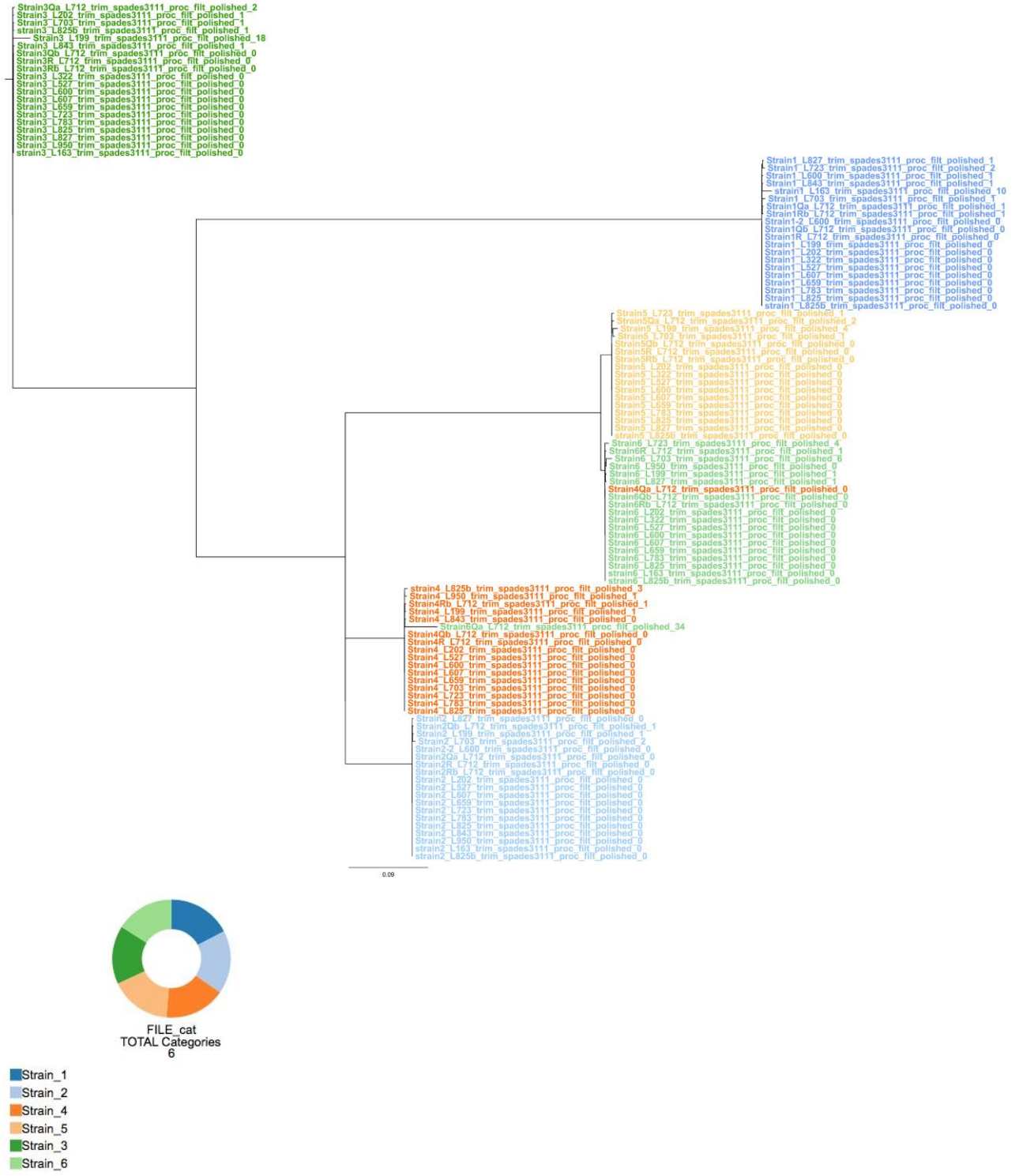
Similarly to what observed with the whole genome SNPs analysis, the main clusters corresponding to the six test strains could be well identified either with or without assembly optimization and again the clusters appeared more homogeneous when optimization was performed. With respect to the whole genome SNPs analysis presented in the previous paragraph, the core genome SNPs analysis resulted in a generally lower intra-cluster variability.

Figure 3. Maximum Parsimony clustering of the core ksnp3 analysis of the whole set of crude assembled contigs (3A) and optimized assemblies (3B).

3A



3B


Strain3Qa_L712_trim_spades3111_proc_filt_polished_2
Strain3_L202_trim_spades3111_proc_filt_polished_1
Strain3_L703_trim_spades3111_proc_filt_polished_1
strain3_L825b_trim_spades3111_proc_filt_polished_1
Strain3_L199_trim_spades3111_proc_filt_polished_18
Strain3_L843_trim_spades3111_proc_filt_polished_1
Strain3Qb_L712_trim_spades3111_proc_filt_polished_0
Strain3R_L712_trim_spades3111_proc_filt_polished_0
Strain3Rb_L712_trim_spades3111_proc_filt_polished_0
Strain3_L322_trim_spades3111_proc_filt_polished_0
Strain3_L527_trim_spades3111_proc_filt_polished_0
Strain3_L600_trim_spades3111_proc_filt_polished_0
Strain3_L607_trim_spades3111_proc_filt_polished_0
Strain3_L659_trim_spades3111_proc_filt_polished_0
Strain3_L723_trim_spades3111_proc_filt_polished_0
Strain3_L783_trim_spades3111_proc_filt_polished_0
Strain3_L825_trim_spades3111_proc_filt_polished_0
Strain3_L827_trim_spades3111_proc_filt_polished_0
Strain3_L950_trim_spades3111_proc_filt_polished_0
strain3_L163_trim_spades3111_proc_filt_polished_0
Strain1_L827_trim_spades3111_proc_filt_polished_1
Strain1_L723_trim_spades3111_proc_filt_polished_2
Strain1_L600_trim_spades3111_proc_filt_polished_1
Strain1_L843_trim_spades3111_proc_filt_polished_1
strain1_L163_trim_spades3111_proc_filt_polished_10
Strain1_L703_trim_spades3111_proc_filt_polished_1
Strain1Qa_L712_trim_spades3111_proc_filt_polished_1
Strain1Rb_L712_trim_spades3111_proc_filt_polished_1
Strain1_L600_trim_spades3111_proc_filt_polished_0
Strain1Qb_L712_trim_spades3111_proc_filt_polished_0
Strain1R_L712_trim_spades3111_proc_filt_polished_0
Strain1_L199_trim_spades3111_proc_filt_polished_0
Strain1_L202_trim_spades3111_proc_filt_polished_0
Strain1_L322_trim_spades3111_proc_filt_polished_0
Strain1_L527_trim_spades3111_proc_filt_polished_0
Strain1_L607_trim_spades3111_proc_filt_polished_0
Strain1_L659_trim_spades3111_proc_filt_polished_0
Strain1_L783_trim_spades3111_proc_filt_polished_0
Strain1_L825_trim_spades3111_proc_filt_polished_0
strain1_L825b_trim_spades3111_proc_filt_polished_0
Strain5_L723_trim_spades3111_proc_filt_polished_1
Strain5Qa_L712_trim_spades3111_proc_filt_polished_2
Strain5_L199_trim_spades3111_proc_filt_polished_4
Strain5_L703_trim_spades3111_proc_filt_polished_1
Strain5Qb_L712_trim_spades3111_proc_filt_polished_0
Strain5R_L712_trim_spades3111_proc_filt_polished_0
Strain5Rb_L712_trim_spades3111_proc_filt_polished_0
Strain5_L202_trim_spades3111_proc_filt_polished_0
Strain5_L322_trim_spades3111_proc_filt_polished_0
Strain5_L527_trim_spades3111_proc_filt_polished_0
Strain5_L600_trim_spades3111_proc_filt_polished_0
Strain5_L607_trim_spades3111_proc_filt_polished_0
Strain5_L659_trim_spades3111_proc_filt_polished_0
Strain5_L783_trim_spades3111_proc_filt_polished_0
Strain5_L825_trim_spades3111_proc_filt_polished_0
Strain5_L827_trim_spades3111_proc_filt_polished_0
strain5_L825b_trim_spades3111_proc_filt_polished_0
Strain6_L723_trim_spades3111_proc_filt_polished_4
Strain6R_L712_trim_spades3111_proc_filt_polished_1
Strain6_L703_trim_spades3111_proc_filt_polished_6
Strain6_L950_trim_spades3111_proc_filt_polished_0
Strain6_L199_trim_spades3111_proc_filt_polished_1
Strain6_L827_trim_spades3111_proc_filt_polished_1
Strain4Qa_L712_trim_spades3111_proc_filt_polished_0
Strain6Qb_L712_trim_spades3111_proc_filt_polished_0
Strain6Rb_L712_trim_spades3111_proc_filt_polished_0
Strain6_L202_trim_spades3111_proc_filt_polished_0
Strain6_L322_trim_spades3111_proc_filt_polished_0
Strain6_L527_trim_spades3111_proc_filt_polished_0
Strain6_L600_trim_spades3111_proc_filt_polished_0
Strain6_L607_trim_spades3111_proc_filt_polished_0
Strain6_L659_trim_spades3111_proc_filt_polished_0
Strain6_L783_trim_spades3111_proc_filt_polished_0
Strain6_L825_trim_spades3111_proc_filt_polished_0
strain6_L163_trim_spades3111_proc_filt_polished_0
strain6_L825b_trim_spades3111_proc_filt_polished_0
strain4_L825b_trim_spades3111_proc_filt_polished_3
Strain4_L950_trim_spades3111_proc_filt_polished_1
Strain4Rb_L712_trim_spades3111_proc_filt_polished_1
Strain4_L199_trim_spades3111_proc_filt_polished_1
Strain4_L843_trim_spades3111_proc_filt_polished_0
Strain6Qa_L712_trim_spades3111_proc_filt_polished_34
Strain4Qb_L712_trim_spades3111_proc_filt_polished_0
Strain4R_L712_trim_spades3111_proc_filt_polished_0
Strain4_L202_trim_spades3111_proc_filt_polished_0
Strain4_L527_trim_spades3111_proc_filt_polished_0
Strain4_L600_trim_spades3111_proc_filt_polished_0
Strain4_L607_trim_spades3111_proc_filt_polished_0
Strain4_L659_trim_spades3111_proc_filt_polished_0
Strain4_L703_trim_spades3111_proc_filt_polished_0
Strain4_L723_trim_spades3111_proc_filt_polished_0
Strain4_L783_trim_spades3111_proc_filt_polished_0
Strain4_L825_trim_spades3111_proc_filt_polished_0
Strain2_L827_trim_spades3111_proc_filt_polished_0
Strain2Qb_L712_trim_spades3111_proc_filt_polished_1
Strain2_L199_trim_spades3111_proc_filt_polished_1
Strain2_L703_trim_spades3111_proc_filt_polished_2
Strain2_L600_trim_spades3111_proc_filt_polished_0
Strain2Qa_L712_trim_spades3111_proc_filt_polished_0
Strain2R_L712_trim_spades3111_proc_filt_polished_0
Strain2Rb_L712_trim_spades3111_proc_filt_polished_0
Strain2_L202_trim_spades3111_proc_filt_polished_0
Strain2_L527_trim_spades3111_proc_filt_polished_0
Strain2_L607_trim_spades3111_proc_filt_polished_0
Strain2_L659_trim_spades3111_proc_filt_polished_0
Strain2_L723_trim_spades3111_proc_filt_polished_0
Strain2_L783_trim_spades3111_proc_filt_polished_0
Strain2_L825_trim_spades3111_proc_filt_polished_0
Strain2_L843_trim_spades3111_proc_filt_polished_0
Strain2_L950_trim_spades3111_proc_filt_polished_0
strain2_L163_trim_spades3111_proc_filt_polished_0
strain2_L825b_trim_spades3111_proc_filt_polished_0
0.09
FILE_cat
TOTAL Categories
6
Strain_1
Strain_2
Strain_4
Strain_5
Strain_3
Strain_6

5.6.3. Whole Genome MLST (wgMLST)

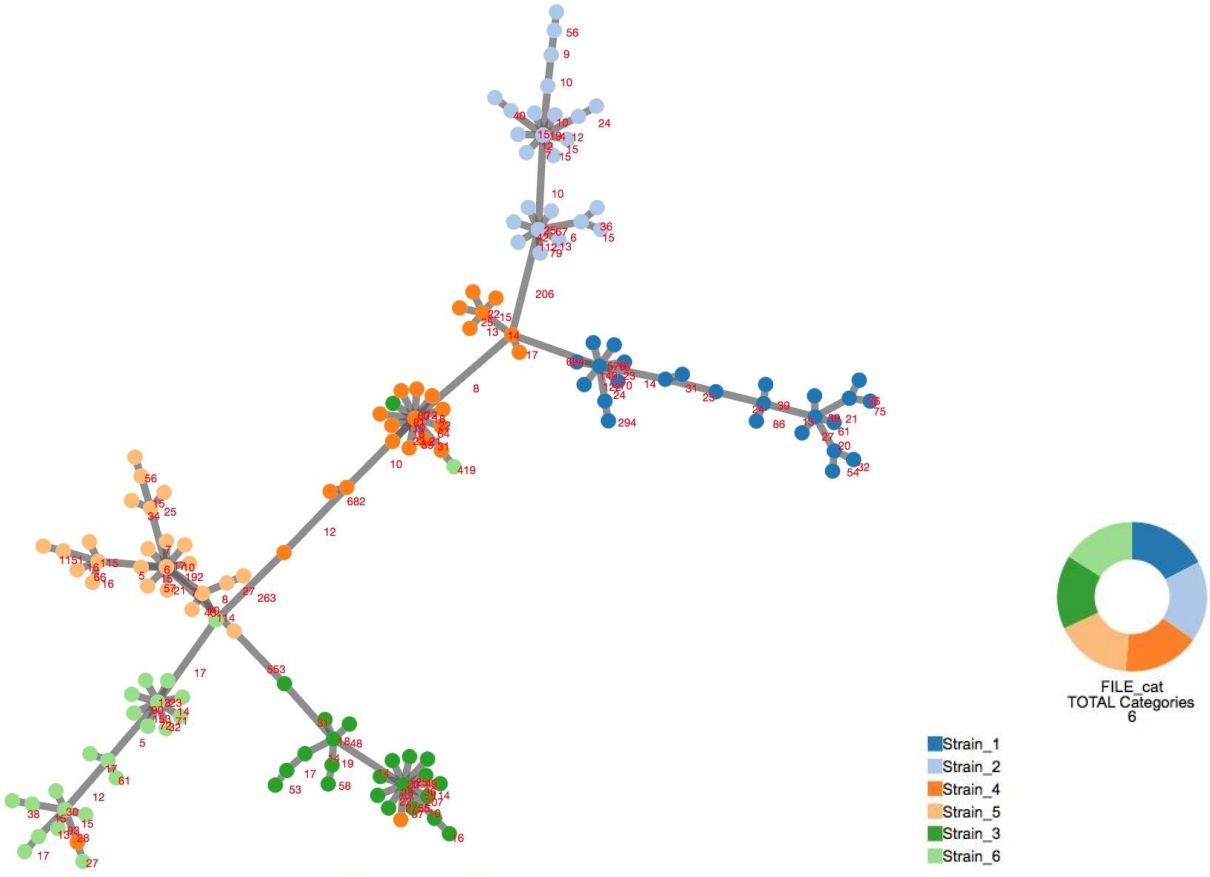
The wgMLST analysis was carried out using the chewBBACA software (Silva *et al*., 2018) installed on the ARIES galaxy instance with a scheme composed of 7601 genes curated by Innuendo EFSA-funded project (http://www.innuendoweb.org/).

The MST resulting from the analysis of the crude assemblies and with the optimized contigs are shown in Figures 4A and 4B, respectively.

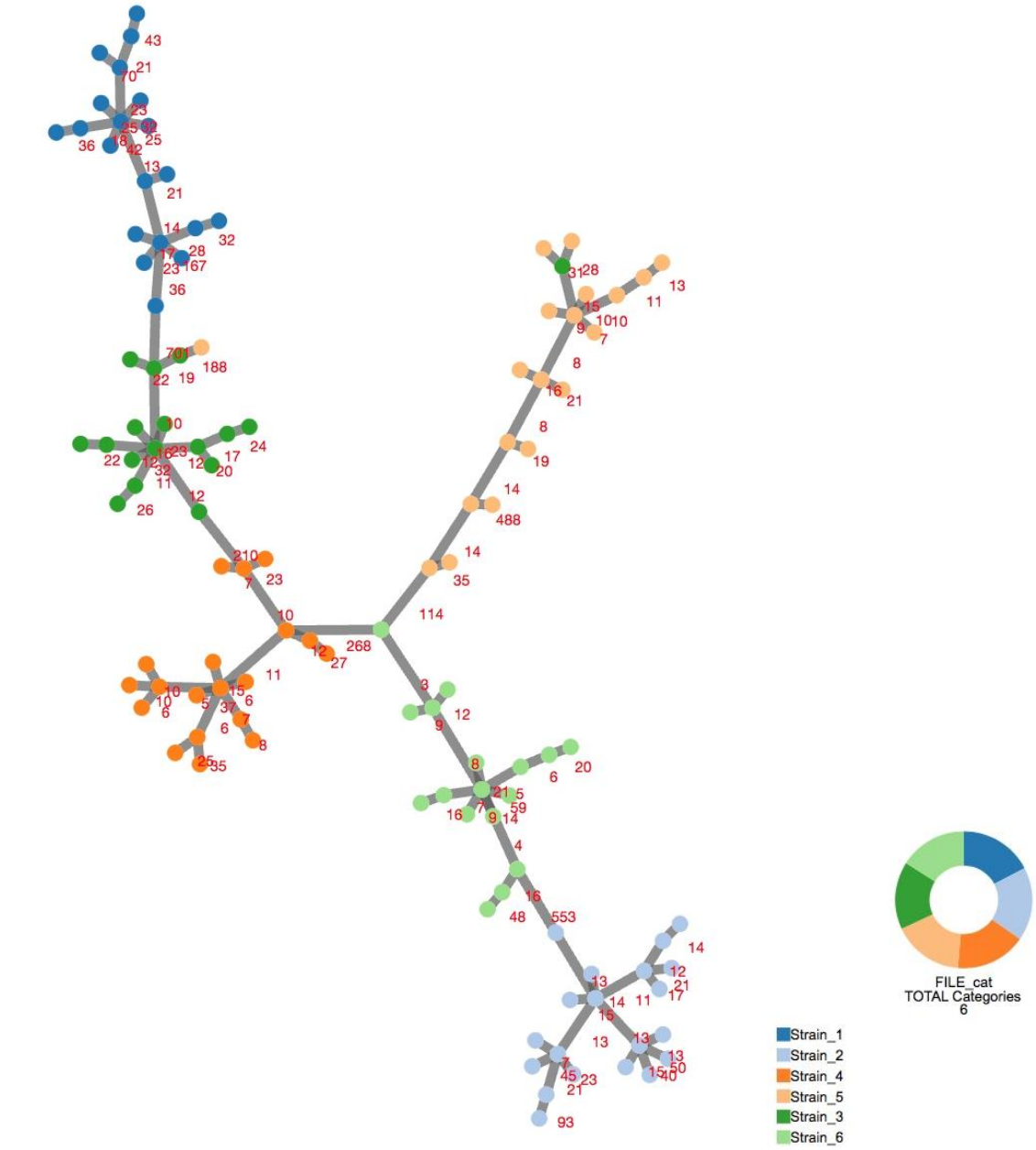
This analysis showed a great variability in terms of allelic differences for the six test strains sequenced in the different participating laboratories, which made the prompt identification of clusters difficult.

Figure 4. Minimum Spanning Tree of the wgMLST allelic profiles of the whole set of crude assembled contigs (4A) and optimized assemblies (4B).

4A



4B


FILE_cat
TOTAL Categories
6
Strain_1
Strain_2
Strain_4
Strain_5
Strain_3
Strain_6

5.6.4. Core Genome MLST on a fixed scheme of genes (cgMLST)

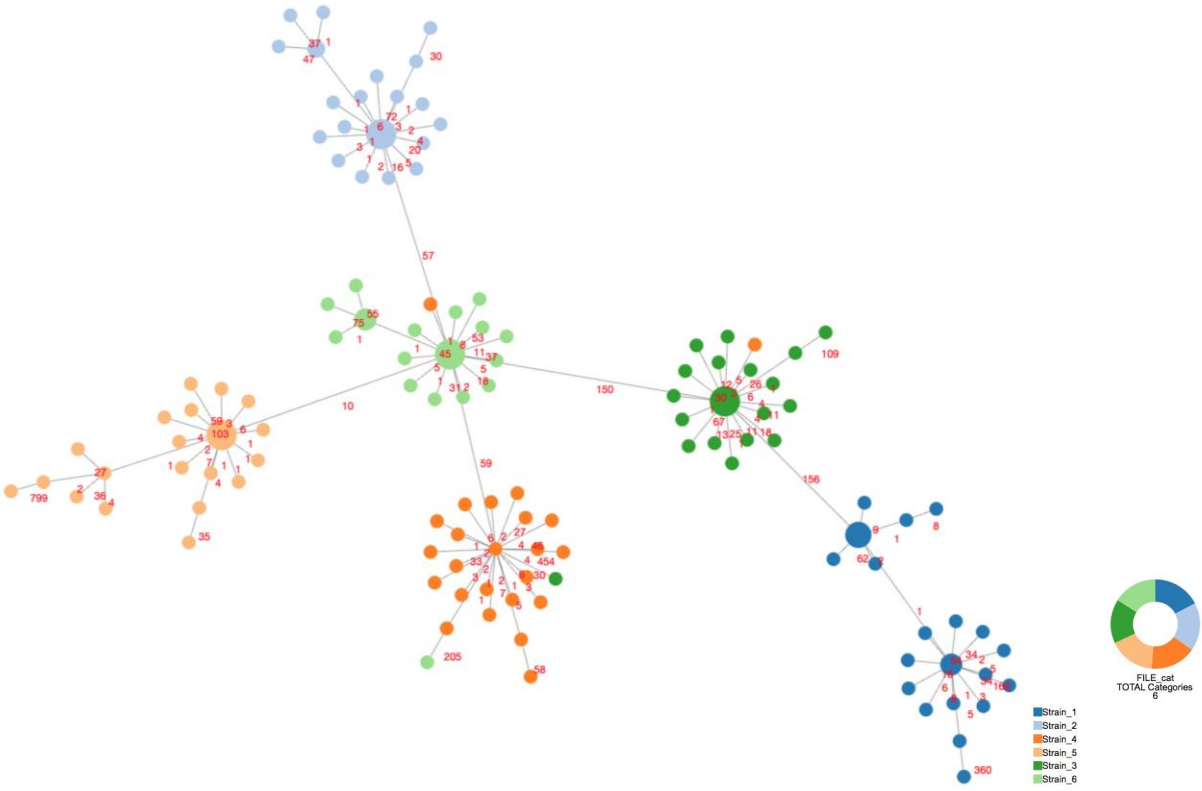
The cgMLST analysis was carried out using the same software used for the wgMLST (Silva *et al*., 2018) but on a fixed scheme of *E. coli* core loci composed of 2360 genes, curated by Innuendo EFSA-funded project (http://www.innuendoweb.org/).

The MST resulting with the crude assembled contigs and with the optimized contigs are shown in Figures 5A and 5B, respectively.

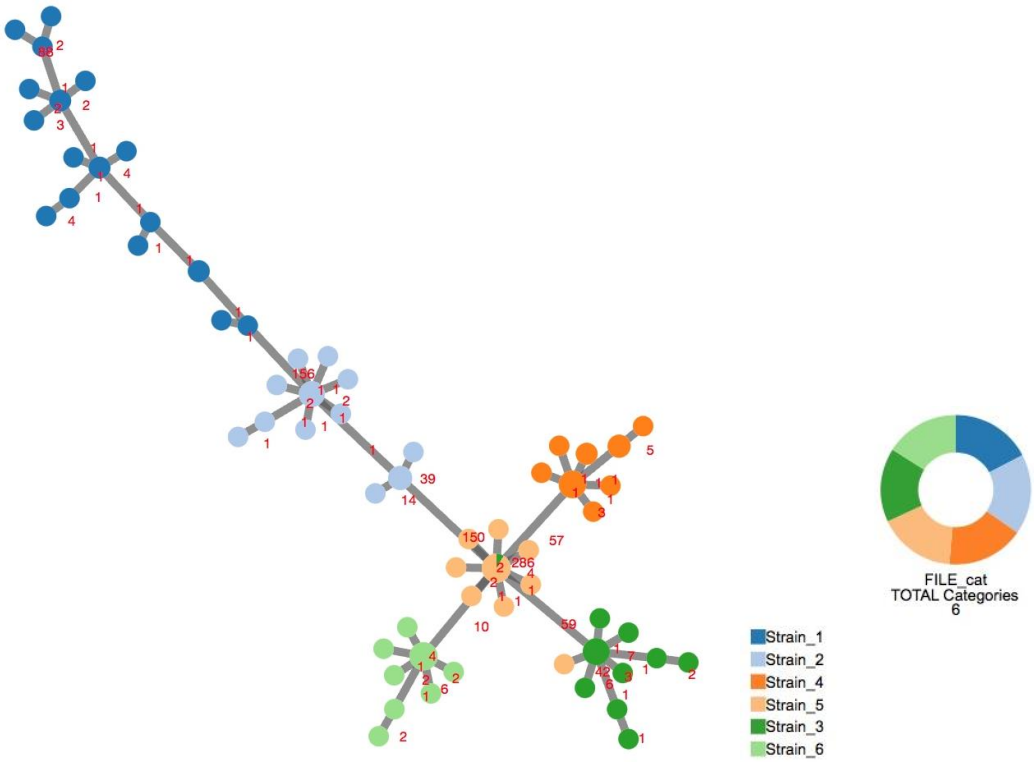
This analysis allowed identifying the six main clusters corresponding to the six test strains. The distances among the sequences (number of allelic differences) of the same test strain produced by the participating labs were much lower when the analysis was performed after assembly optimization.

Figure 5. Minimum Spanning Tree of the fixed cgMLST scheme allelic profiles of the whole set of crude assembled contigs (A) and optimized assemblies (B).

5A



5B


Strain_1
Strain_2
Strain_4
Strain_5
Strain_3
Strain_6
FILE_cat
TOTAL Categories
6

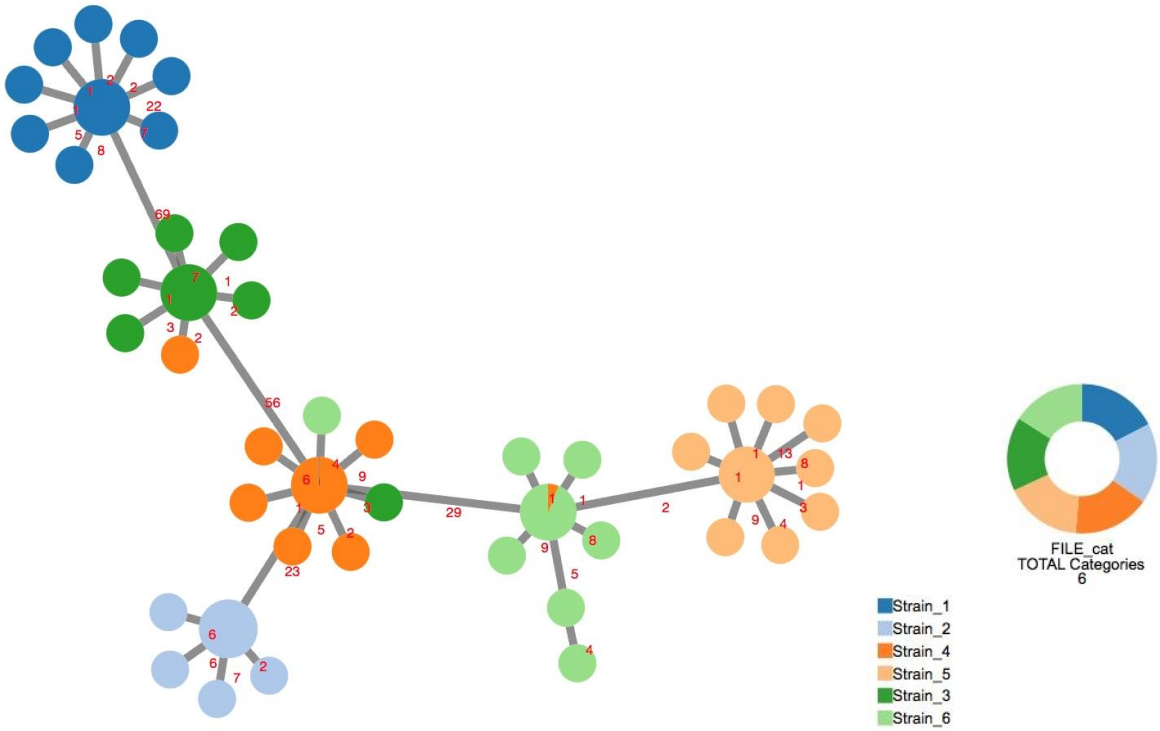
5.6.5. Core Genome MLST on a dynamic scheme of shared genes

The table of loci obtained with the wgMLST analysis (paragraph 5.6.3) was used to extract the allelic profiles of the loci shared among all (100 %) the test sequences (dynamic core) with the chewBBACA software. The analysis was repeated with the crude assembled contigs and optimized assemblies obtaining the following goeBURST sizes: 1152 loci (crude assemblies) and 2546 loci (optimized assemblies).

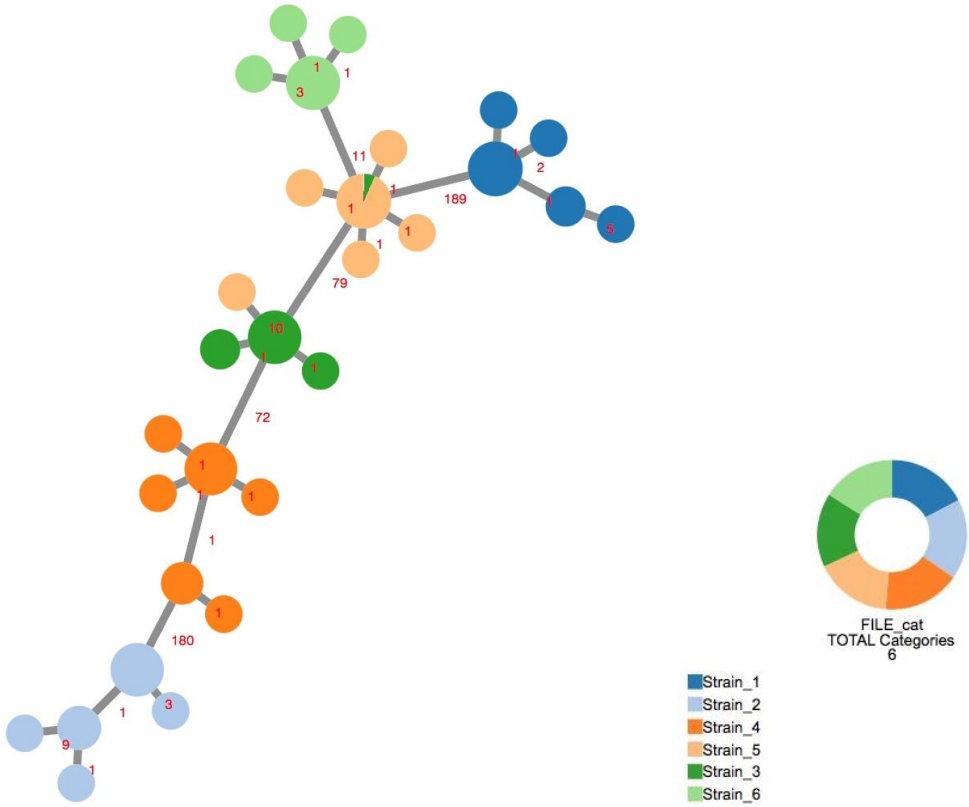
The respective MST are shown in Figures 6A and 6B. The MST topology allowed in both cases to identify the six main clusters corresponding to the six test strains and the distances among the sequences of the same test strain produced by the participating labs were much lower than in the previously described MSTs, particularly when the analysis was performed after assembly optimization.

Figure 6. Minimum Spanning Tree of the dynamic cgMLST scheme allelic profiles of the whole set of crude assembled contigs (6A) and optimized assemblies (6B).

6A



6B


1
1
3
11
1
2
189
1
1
1
79
10
1
1
72
1
1
1
1
180
3
1
9
1
FILE_cat
TOTAL Categories
6
Strain_1
Strain_2
Strain_4
Strain_5
Strain_3
Strain_6

6. Concluding remarks

Whole genome sequencing has become a realistic alternative to conventional molecular typing methods for bacterial isolates, including Shiga toxin-producing *E. coli*. The analysis of the sequences provided by the laboratories participating in this first voluntary inter-laboratory exercise on WGS induces the following remarks:

1. A good participation rate was observed: 21 Laboratories including 18 NRLs (51.4 % of the whole *E. coli* network) and 3 Italian OLs participated in the study, confirming the interest in this emerging technology.
2. The results of the study highlighted that the majority of the participating laboratories produce whole genome sequences in their routine workflow that can be used to correctly characterize the sequenced strains, although the laboratories were not provided with sequencing specifications (*e.g*. depth of sequencing or average quality etc.).
3. The sequences produced across the network are nevertheless highly heterogeneous (highly different values of depth calculated on the seven genes of the MLST scheme, assembly coverage and N50). High variability is often detected also among sequences produced within the same lab, suggesting non-standardized in-lab workflows.
4. The MLST type of the vast majority of WGS could be correctly identified, with only 15/144 sequences which did not return the expected ST.
5. EURL-VTEC WGS PT pipeline – MLST seven genes - validation: 89.6 % of the WGS produced in non-standardized workflows correctly typed with no MLST types misassigned, 89.6 % sensitivity and 100 % specificity.
6. The detection of the H11 flagellar antigen coding gene was not affected by the large differences in the quality of the sequences. As a matter of fact, all the test strains could be correctly typed as H11 by using all the sequences analysed.
7. The identification of the correct serogroup O26 was also possible for the vast majority of the test sequences. Only six out of 144 WGS were not typed (4.2 %).
8. EURL-VTEC WGS PT pipeline – serotyping - validation: 95.8 % sensitivity, 100 % specificity.
9. The detection of the main virulence genes by applying the EURL-VTEC WGS PT pipeline was possible from the majority of the received sequences. Excluding the errors associated with the inversion of test strains, the presence of *stx2* could not be detected only from a total of four sequences of test strains 5 and 6 (2.8 %), while the presence of the genes *ehxA*, *stx1* and *eae* could always be correctly identified.

10. EURL-VTEC WGS PT pipeline – basic virulotyping - validation: 97.2 % sensitivity and 100 % specificity for *stx2* gene; for the other genes (*stx1, eae* and *ehxA*) both sensitivity and specificity 100 %.
11. SNP analysis was always effective in detecting the right clusters and the inversion of the test strains regardless the optimization of the assemblies. The highest readibility of the dendrogram was assured by the core SNPs analysis.
12. cg/wgMLST provided a good resolution of the clusters only with the cgMLST, while the dispersion of the clusters obtained with the wgMLST analysis did not allow a prompt recognition of the groups either using optimized or non-optimized contigs.
13. The most readable phylogenetic analysis based on MLST of the dataset was obtained with the dynamic cgMLST with both otimezed and non-optimezed assemblies and was always effective in detecting the right clusters and the inversion of the test strains. Howerver, the optimization of the assembly process produces contigs whose analysis allowed the identification of the largest number of shared loci (goeBURST), thus increasing the sensitivity.

6. Bibliography

- Gardner, S. N., Slezak, T. & Hall, B. G. 2015. kSNP3.0: SNP detection and phylogenetic analysis of genomes without genome alignment or reference genome. *Bioinformatics,* 31**,** 2877-8.
- Matthews, T. C., Bristow, F. R., Griffiths, E. J., Petkau, A., Adam, J., Dooley, D., Kruczkiewicz, P., Curatcha, J., Cabral, J., Fornika, D., Winsor, G. L., Courtot, M., Bertelli, C., Roudgar, A., Feijao, P., Mabon, P., Enns, E., Thiessen, J., Keddy, A., Isaac-Renton, J., Gardy, J. L., Tang, P., Carrico, J. A., Chindelevitch, L., Chauve, C., Graham, M. R., Mcarthur, A. G., Taboada, E. N., Beiko, R. G., Brinkman, F. S., Hsiao, W. W. & Van Domselaar, G. 2018. The Integrated Rapid Infectious Disease Analysis (IRIDA) Platform. *bioRxiv*.
- Nascimento, M., Sousa, A., Ramirez, M., Francisco, A. P., Carrico, J. A. & Vaz, C. 2017. PHYLOViZ 2.0: providing scalable data integration and visualization for multiple phylogenetic inference methods. *Bioinformatics,* 33**,** 128-129.
- Silva, M., Machado, M. P., Silva, D. N., Rossi, M., Moran-Gilad, J., Santos, S., Ramirez, M. & Carrico, J. A. 2018. chewBBACA: A complete suite for gene-by-gene schema creation and strain identification. *Microb Genom*.
- Wirth, T., Falush, D., Lan, R., Colles, F., Mensa, P., Wieler, L. H., Karch, H., Reeves, P. R., Maiden, M. C., Ochman, H. & Achtman, M. 2006. Sex and virulence in *Escherichia coli*: an evolutionary perspective. *Mol Microbiol,* 60**,** 1136-51.

Annex 1.

Description of the analytical flow of the "EURL-VTEC WGS PT pipeline" (Galaxy Version 1.0)

The EURL-VTEC WGS PT pipeline was operated through ARIES webserver (https://www.iss.it/site/aries). It was developed to allow the automatic analysis of whole genome sequences of *E. coli*, by using in input the raw data obtained from the sequencing platform in ".fastq" format and producing in output a detailed report of typing features identified in the analysed sequence.

The details of the analytical flow are listed below.

- **Raw data quality check**

  A preliminary quality check step was operated by applying FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) on the raw .fastq data received, either paired-end or single-end.

- **Trimming**

  The tool "FASTQ positional and quality trimming" (Cuccuru et al., 2014) was used to remove the adaptors and discard low quality regions.

  The following parameters were applied:

  - single-end reads: Maximum length trimming 360
    - Left-side trimming 10
    - Right-side trimming 0
    - Minimum Phred quality score for right-side trimming 25
    - Average Phred quality score for right-side trimming 27
    - Minimum length filtering 50
  - paired-end reads: Maximum length trimming 300
    - Left-side trimming 17
    - Right-side trimming 0
    - Minimum Phred quality score for right-side trimming 25
    - Average Phred quality score for right-side trimming 27
    - Minimum length filtering -1

- **Virulotyping**

  The tool “patho_typing” developed by the INNUENDO Project was used to obtain a list of the pathotypes present in the reads by mapping them to the sequences of the *E. coli* virulence genes database curated by the Statens Serum Institut (SSI) & the Danmarks Tekniske Universitet (DTU) (Joensen *et al.*, 2014) setting the following parameters:
  Minimum Gene Coverage: 90
  Minimum Gene Identity: 90
  Minimum Gene Depth: 15

- **Multi Locus Sequence Typing**

  The trimmed reads were used to identify the Sequence Type by using SRST2 tool (Inouye *et al*., 2014) and the MLST scheme based on seven housekeeping loci of *E. coli* (Wirth *et al.*, 2006).

- **Assembly (A5 for Illumina and SPAdes for IonTorrent data)**

  The trimmed reads were used for assembling contigs with the tool “A5 pipeline” (Tritt *et al*., 2012) for paired-end reads and the tool SPAdes 3.11.1 (Bankevich *et al*., 2012) for single-end reads, with default parameters. The contigs assembled with SPAdes were then subjected to filtering with the tool “Filter SPAdes output”, part of SPAdes suite, with the following parameters:
  Length cut-off: 1000
  Coverage cut-off: 10

- **Assembly statistics**

  The tool “Check bacterial contigs” (Cuccuru *et al.*, 2014) was used to calculate the N50 value for the assembled contigs.

- **Serotyping**

  The serotype was obtained through application of the blastn software (Cock *et al*., 2015) to the assembled contigs, by comparing the sequences to those of the O and H serotype databases curated by SSI & DTU (Joensen *et al*., 2015) setting the following parameters:
  Expect value (E) for saving hits: 0.001
  Query strands: both

Filter query sequence with dust: yes

Number of aligned sequences to keep: 10

Percent identity cutoff: 95.0

Assembly optimization pipeline

Assembly optimization was kindly performed by Dr. Joao Andrè Carrico through the use of INNUca pipeline according to the instructions provided in the User Manual available online (https://github.com/B-UMMI/INNUca) with default parameters. INNUca pipeline only accepted paired-end reads in input and for this reason all single-end reads were not subjected to assembly optimization and were not used for the phylogenetic analyses, regardless their quality. Thus, only the assemblies deriving from paired-end reads and satisfying the preliminary quality check were optimized and used for the corresponding phylogenetic analyses.

Description of the tools used for phylogenetic comparison

The ksnp3 tool was operated through ARIES platform according to the user manual (Gardner *et al*., 2015). In detail, the kmer size was selected for each set of sequences analysed, using the kchooser tool included in the ksnp3 pipeline, which selects the optimum value as that producing the highest number of unique kmers of the median length in all the genomes of the dataset.

The core genome SNPs analysis was performed by selecting “Yes” for the “Calculate core SNPs and core parsimony tree” option and by using 0.90 as the “Minimum fraction of genomes with locus” parameter.

The dendrograms were obtained by using the maximum parsimony algorithm and the dendrograms were visualized and coloured by using FigTree software v 1.4.3 (http://tree.bio.ed.ac.uk/software/figtree/).

The wgMLST, cgMLST and dynamic cgMLST analyses were performed by using chewBBACA (Silva *et al.*, 2018) tool through ARIES webserver. In detail, for wgMLST and cgMLST the allele call was performed on the fixed scheme of 7601 and 2360 genes, respectively, curated by INNUENDO EFSA-funded project (http://www.innuendoweb.org/) and default parameters. The allele call for the dynamic cgMLST analysis was instead executed with the “ExtractCgMLST” function operated on the matrix of alleles obtained when performing wgMLST, using 0.90 as “maximum presence” value.

The Minimum Spanning Trees obtained with the MLST-based analyses were visualized and coloured using Phyloviz online 2.0 (Nascimento *et al*., 2017).

Either the two ksnp3 and the three MLST-based analyses were performed before and after assembly optimization.

Bibliography

- Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., Lesin, V. M., Nikolenko, S. I., Pham, S., Prjibelski, A. D., Pyshkin, A. V., Sirotkin, A. V., Vyahhi, N., Tesler, G., Alekseyev, M. A. & Pevzner, P. A. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol,* 19**,** 455-77.
- Cock, P. J., Chilton, J. M., Gruning, B., Johnson, J. E. & Soranzo, N. 2015. NCBI BLAST+ integrated into Galaxy. *Gigascience,* 4**,** 39.
- Cuccuru, G., Orsini, M., Pinna, A., Sbardellati, A., Soranzo, N., Travaglione, A., Uva, P., Zanetti, G. & Fotia, G. 2014. Orione, a web-based framework for NGS analysis in microbiology. *Bioinformatics,* 30**,** 1928-9.
- Gardner, S. N., Slezak, T. & Hall, B. G. 2015. kSNP3.0: SNP detection and phylogenetic analysis of genomes without genome alignment or reference genome. *Bioinformatics,* 31**,** 2877-8.
- Inouye, M., Dashnow, H., Raven, L. A., Schultz, M. B., Pope, B. J., Tomita, T., Zobel, J. & Holt, K. E. 2014. SRST2: Rapid genomic surveillance for public health and hospital microbiology labs. *Genome Med,* 6**,** 90.
- Joensen, K. G., Scheutz, F., Lund, O., Hasman, H., Kaas, R. S., Nielsen, E. M. & Aarestrup, F. M. 2014. Real-time whole-genome sequencing for routine typing, surveillance, and outbreak detection of verotoxigenic *Escherichia coli*. *J Clin Microbiol,* 52**,** 1501-10.
- Joensen, K. G., Tetzschner, A. M., Iguchi, A., Aarestrup, F. M. & Scheutz, F. 2015. Rapid and Easy In Silico Serotyping of *Escherichia coli* Isolates by Use of Whole-Genome Sequencing Data. *J Clin Microbiol,* 53**,** 2410-26.
- Matthews, T. C., Bristow, F. R., Griffiths, E. J., Petkau, A., Adam, J., Dooley, D., Kruczkiewicz, P., Curatcha, J., Cabral, J., Fornika, D., Winsor, G. L., Courtot, M., Bertelli, C., Roudgar, A., Feijao, P., Mabon, P., Enns, E., Thiessen, J., Keddy, A., Isaac-Renton, J., Gardy, J. L., Tang, P., Carrico, J. A., Chindelevitch, L., Chauve, C., Graham, M. R., Mcarthur, A. G., Taboada, E. N., Beiko, R. G., Brinkman, F. S., Hsiao, W. W. & Van Domselaar, G. 2018. The Integrated Rapid Infectious Disease Analysis (IRIDA) Platform. *bioRxiv*.
- Nascimento, M., Sousa, A., Ramirez, M., Francisco, A. P., Carrico, J. A. & Vaz, C. 2017. PHYLOViZ 2.0: providing scalable data integration and visualization for multiple phylogenetic inference methods. *Bioinformatics,* 33**,** 128-129.
- Silva, M., Machado, M. P., Silva, D. N., Rossi, M., Moran-Gilad, J., Santos, S., Ramirez, M. & Carrico, J. A. 2018. chewBBACA: A complete suite for gene-by-gene schema creation and strain identification. *Microb Genom*.
- Tritt, A., Eisen, J. A., Facciotti, M. T. & Darling, A. E. 2012. An integrated pipeline for *de novo* assembly of microbial genomes. *PLoS One,* 7**,** e42304.
- Wirth, T., Falush, D., Lan, R., Colles, F., Mensa, P., Wieler, L. H., Karch, H., Reeves, P. R., Maiden, M. C., Ochman, H. & Achtman, M. 2006. Sex and virulence in *Escherichia coli*: an evolutionary perspective. *Mol Microbiol,* 60**,** 1136-51.