EUCHINASAFE
中欧食品安全

Deliverable number: D4.7
Deliverable title: Report on the application of techniques for sequencing and WGS analysis

Work Package No.:	WP 4
Lead beneficiary:	Teagasc
Due date (project month - dd/mm/yyyy):	M42
Actual delivery date (project month - dd/mm/yyyy):	M60 – 19/08/2022

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 727864 and from the Chinese Ministry of Science and Technology (MOST).

Delivering an Effective, Resilient and Sustainable EU-China Food Safety Partnership

Grant Agreement number:
727864 — EU-China-Safe

Acknowledgements

This report forms part of the deliverables of the project “EU-China-Safe” which has received funding from the European Union’s Horizon 2020 Research and Innovation programme under Grant Agreement No 727864 and from the Chinese Ministry of Science and Technology (MOST).

EU-China-Safe aims at reducing food fraud and improving food safety through focusing on improving food legislation, food inspection and increasing access to information across Europe and China. State-of-the-art technologies including a virtual laboratory will create a unique space to share and demonstrate best practice. The use of innovative technologies will result in improved detection of adulteration of food products as well as increased traceability and transparency of global supply chains.

The project runs from September 2017 to August 2021. It involves 33 partners and is coordinated by QUB (The Queen’s University of Belfast, UK).

More information on the project can be found at www.euchinasafe.eu (website in construction)

The content of this report does not reflect the official opinion of the European Commission and/or Chinese government . Responsibility for the information and views expressed therein lies entirely with the author(s).

Document control page:

Deliverable Title	**Report on the harmonisation of laboratory and *in silico* techniques for whole genome sequencing (WGS) of bacteria and SARS-CoV-2 virus and bioinformatic analysis**
Author (writer/editor and short name of the organisation)	Professor Séamus Fanning
Contributors (co-authors and short names of the organisations)	Dr Guerrino Macori
Version number (VX.Y)	V3
Version date (dd/mm/yyyy)	19/08/2022
Last modified by (person and organisation name)	Guerrino Macori
Rights (e.g. NA, Intellectual Property Rights, copyright, …)	Unpublished research not for widespread dissemination until paper is published (keep internal).

Revision history:

Version	**Date**	**Modified by**	**Comments**

Nature of the deliverable

ORDP	Open Research Data Pilot	
R	Document, report (excluding the periodic and final reports)	R
DEM	Demonstrator, pilot, prototype, plan designs	
DEC	Websites, patents filing, press & media actions, videos, etc.	
E	Ethics	
OTHER	Software, technical diagram, etc.	

Dissemination Level

PU	Public, fully open, e.g. web	
CO	Confidential, restricted under conditions set out in Model Grant Agreement	CO
CI	Classified, information as referred to in Commission Decision 2001/844/EC	

Table of contents

1. SUMMARY

A harmonised SOP dedicated to the application of a bioinformatics pipeline was developed to enable comparable methods applying this approach to data generated by whole genome sequencing (WGS) between partners. An inter-laboratory ring trial was carried out as part of this task to test these bioinformatic pipeline methods, among all partners. To this end twenty-four bacterial isolates from the UCD-Centre for Food Safety strain collection were sent (blinded) to the partners. WGS of these isolates was carried out using technical methods developed in ***D4.8***, and then a comparison of the performance of each laboratory, described in this deliverable.

In addition, as an extra item, arising from the covid-19 pandemic and the necessity to protect the security of the food chain, protocols focused on the sequencing of this virus, SARS-CoV-2 of importance to human health, were also included.

This deliverable is focused only on the application and implementation these bioinformatic-based techniques.

2. INTRODUCTION

Whole Genome Sequencing (WGS) is a molecular technique whereby purified DNA from a bacterium of interest is fragmented and undergoes a protocol known as library preparation. The prepared DNA is then sequenced on a dedicated platform which generates an output of genomic sequences. The sequencing data generated can be used to identify single nucleotide differences (single nucleotide polymorphisms or SNPs) which describe allelic mutations, allowing the differentiation between genomes of organisms, or simply providing a deeper understanding of their genetic makeup (Land et al., 2015) and the possible artificial differences introduced during sample preparation or bioinformatics analysis. The application of WGS can provide a range of information, including the presence of mobile genetic elements (MGE e.g., plasmids/bacteriophages), virulence factors such as toxin-encoding genes, or may be used to examine DNA modifications and methylation profiles of foodborne pathogens.

WGS may be carried out on a number of platforms including Illumina and Oxford Nanopore Technology. Illumina is the preferred platform for short reads, providing quick turnaround of high-quality results at a relatively low cost (Mitchell et al., 2021), while Oxford Nanopore Technology sequencing, on devices such as MinION, are used for single molecule long read sequencing. While longer sequencing read lengths may help to resolve repetitive DNA repeats and detect epigenetic markers, this technology requires more template DNA and has higher error rates (Quail et al., 2012).

A number of bioinformatic tools allow the analysis of the data output which may provide important insights in terms of outbreak surveillance, forensics, metabolic modelling and metagenome analysis (Land et al., 2015). These tools facilitate genome assembly and subsequent comparative analysis, detection of virulence and AMR genes, SNP calling for genetic comparison between bacterial isolates, and phylogenetic analyses. Following sequencing, steps are taken to decipher the data output to enable the user to draw meaningful conclusions. The general pipeline applied to sequencing output includes quality checks of raw reads (e.g., FastQC), adapter trimming (e.g., Porechop, Trimmomatic), *de novo* genome assembly (e.g., SPAdes), followed by quality assessment of the generated assemblies (e.g., Quast). The draft genome obtained with the assemblies are then annotated (e.g., Prokka) (Mitchell et al., 2021). Using databases available over the internet (e.g., ResFinder, PlasmidFinder, Virulence Finder Database), and bioinformatics tools (e.g., ABRicate, SeqSero, PubMLST), AMR-encoding genes, plasmids and virulence genes may be identified within the genome (Macori et al., 2021). This information can offer insight into the resistance phenotypes which may be expressed, as well as the level of virulence expected (Bogaerts et al., 2019; Wyres et al., 2014). In two different laboratories of the partners participating in this exercise, genomic DNA was extracted from fresh cultures and prepared for sequencing. The raw data generated were analysed for the assessment of the quality and the correct identification of molecular markers, including typing genes, AMR-genes and presence of genomic features such as plasmids. To evaluate the correct generation of the outputs, a detailed study of SNPs detection was carried out at different level, including raw reads, cleaned reads and assembled genomes.

The aim of this project was to develop and implement a shared vision of best practice within the EU and China in an effort to enhance food safety, deter food fraud, deliver mutual recognition of data and standards and support the flow of agri-food trade between the two trading blocks in a way that better protects the consumer. In this work, we report on the techniques for sequencing and WGS analysis. This involved:

1. Selecting a set of bacterial isolates for sequencing in partner laboratories.
2. Application of Illumina MiSeq techniques to generate bacterial genome sequences and
3. Downstream WGS analysis using a bioinformatics pipeline.

Note- this deliverable should be considered alongside D4.8, which provides the details of the associated SOPs.

3. APPLICATION OF TECHNIQUES FOR SEQUENCING

3.1. BACTERIAL CULTURE, GENOMIC DNA EXTRACTION AND QUANTIFICATION

3.1.1. Culture

Twenty-four selected bacteria, representing three genera of importance to food safety were included for this study. Their details are shown in **Table 1.**

Table 1. Details on the numbers and codes of the isolates distributed to the partner of the project.

number isolate	Strain	Species	number isolate	Strain	Species	number isolate	Strain	Species
1	CFS3535	*1*	9	CFS4391	*2*	17	F2151	*3*
2	CFS3536	*1*	10	CFS4392	*2*	18	F2152	*3*
3	CFS3537	*1*	11	CFS4393	*2*	19	F2153	*3*
4	CFS3538	*1*	12	CFS4394	*2*	20	F2154	*3*
5	CFS3539	*1*	13	CFS4395	*2*	21	F2155	*3*
6	CFS3540	*1*	14	CFS4396	*2*	22	F2160	*3*
7	CFS3541	*1*	15	CFS4397	*2*	23	F2161	*3*
8	CFS3542	*1*	16	CFS4398	*2*	24	F2166	*3*

The isolates were cultured as described in the SOP, provided in ***D4.8, section 3.1.1.***

3.1.2. DNA Extraction

Bacterial genomic DNA (gDNA) was purified from 24 selected bacteria, representing three genera of importance to food safety. The details of the extraction method are provided in an SOP, described in ***D4.8, section 3.1.2.***

3.1.3. Quantification of bacterial genomic DNA

Quantification of bacterial genomic is described in a detailed SOP in ***D4.8, section 3.1.3.***

3.2 LIBRARY PREPARATION

3.2.6 Fragmentation/End Prep

DNA fragmentation was carried out using the NEBNext® Ultra™ II FS DNA Library Prep Kit for Illumina®. The details of this protocol are described in an SOP shown in ***D4.8, section 3.2.6.***

3.2.7 Adaptor Ligation

Adaptor ligation and the steps involved are described in an SOP shown in ***D4.8, section 3.2.7.***

3.2.8 Size Selection of Adaptor-ligated DNA fragment Sizes >550 bp

Once adaptors are attached, these are then size selected, as described in an SOP shown in ***D4.8, section 3.2.8.***

3.2.9 PCR Enrichment of Adaptor-Ligated DNA

Prior to a clean-up step, the ligated adaptors are then subjected to PCR enrichment to add “indexes” to the samples being tested. This step described in an SOP shown in ***D4.8, section 3.2.9.***

3.2.10 Clean-up of PCR Reaction

Clean-up of the Adaptor ligated DNA is described in an SOP shown in ***D4.8, section 3.2.10.***

3.2.11 Normalisation

Finally, in order to normalise each of the samples within the library, two measurements for each sample are needed - fragment size and DNA concentration. Quantification of bacterial genomic DNA may be carried out using the Qubit™ 2.0 Fluorometer in combination with the Qubit™ dsDNA HS (High Sensitivity) Assay Kit, or alternatively using the Nanodrop™ 1000 Fluorospectrometer, while fragment size is assessed using the Agilent Tapestation This is described in an SOP shown in ***D4.8, section 3.2.11.***

3.2.12 Library Denaturation

After the clean-up step above, and prior to loading, the library is denatured described in an SOP shown in ***D4.8, section 3.2.12.***

3.2.13 MiSeq Loading

For loading samples onto the Illumina MiSeq platform, see the SOP shown in ***D4.8, section 3.2.13.***

4. OUTPUT FROM WGD

The Bioinformatics Pipeline used is detailed in full in the appendices of ***D4.8***. However, the general scheme involved initial quality control (FastQC, MultiQC), followed by filtering and trimming of reads (fastP). Quality Control was then repeated and subsequently reads filtering and trimming were repeated as necessary. K-mer analysis against known genome databases was performed for species identification. *De novo* assembly using SPADES was carried out. File ordering was carried out to allow for collection and renaming of scaffolds. Contigs were examined and counted, and contigs smaller than 500 bp were removed using a script in perl language optimised at CFS. Following this, the quality of results were assessed using the Quality Assessment Tool for Genome Assemblies (QUAST). Genome Size estimation was carried out using K-mer analysis. Annotation using Prokka was carried out, followed by MLST (mlst + PubMLST), Resistome analysis (ABRicate + Resfinder, Argannot, CARD, NCBI), Virulome analysis (ABRicate + VFDB) and Variant identification (Snippy). Other Antimicrobial Resistance Gene Databases such as ARDB, BacMet, CARD, etc were then searched, followed by Plasmid sequence detection with Plasmid Finder and for *Salmonella* sp. isolates, SeqSero was used for the prediction of the serotypes. Finally, the GO annotation was prepared. A schematic representation is shown in **Figure-1.**

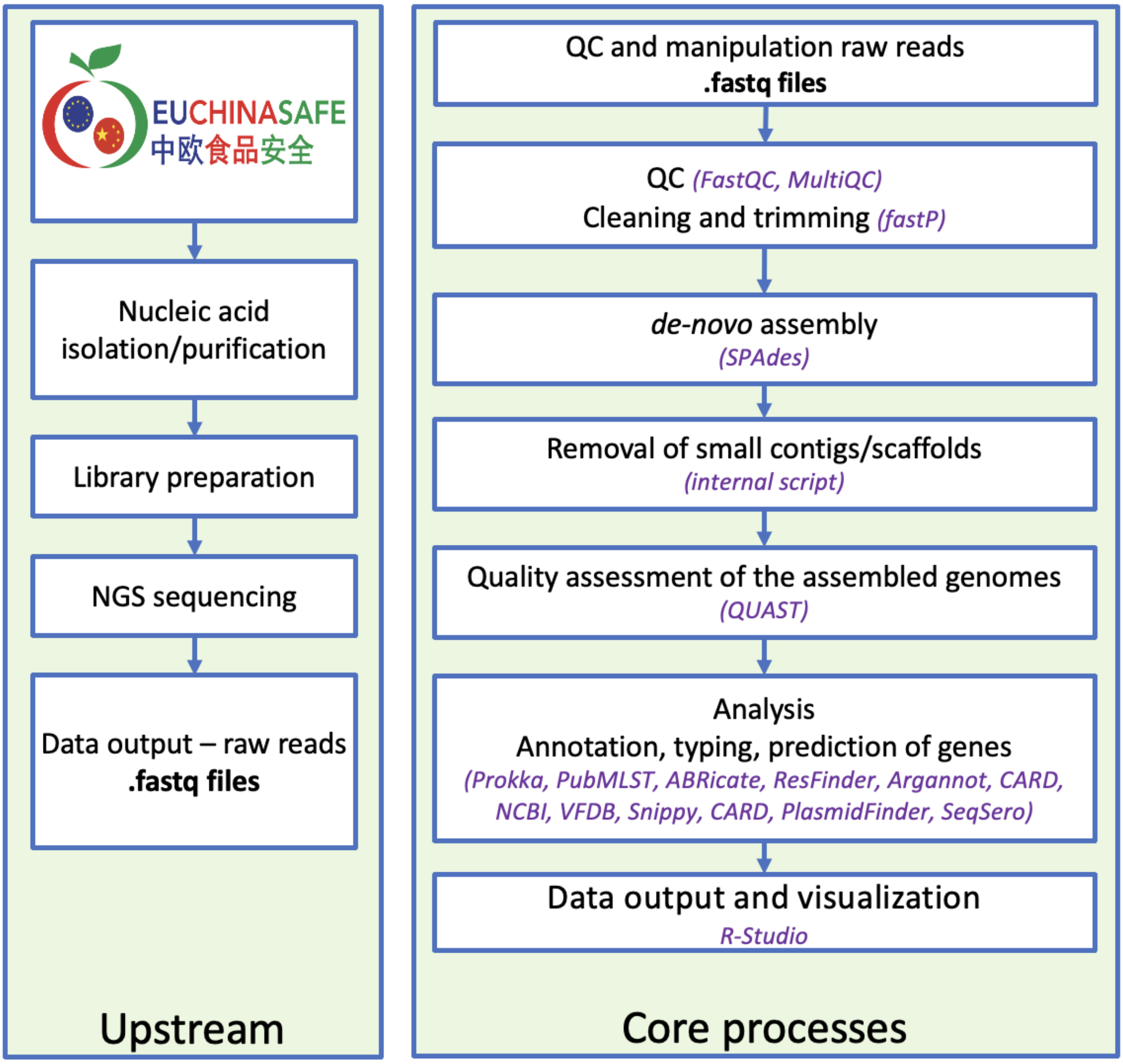
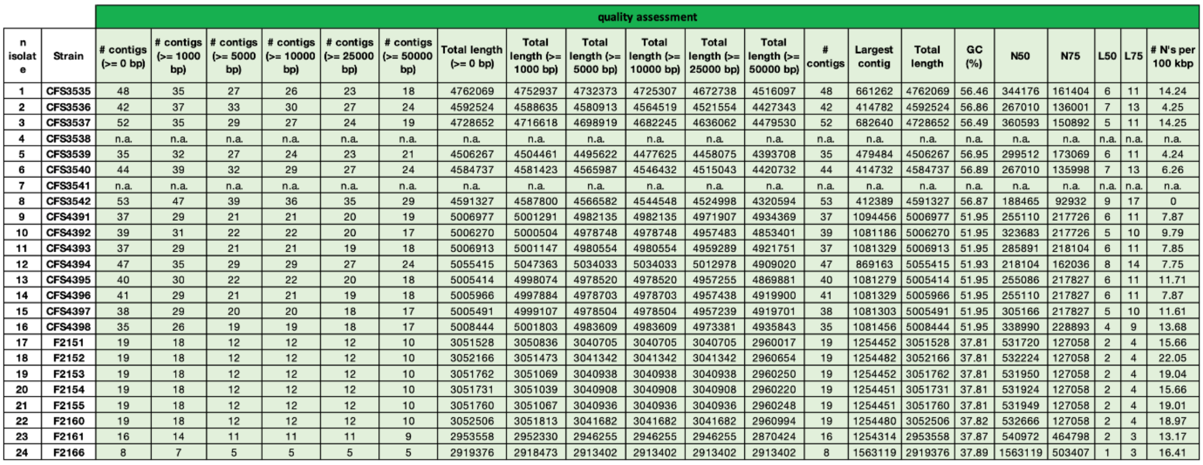

EUCHINASAFE
中欧食品安全
Nucleic acid isolation/purification
Library preparation
NGS sequencing
Data output – raw reads .fastq files
Upstream
QC and manipulation raw reads .fastq files
QC (FastQC, MultiQC) Cleaning and trimming (fastP)
de-novo assembly (SPAdes)
Removal of small contigs/scaffolds (internal script)
Quality assessment of the assembled genomes (QUAST)
Analysis Annotation, typing, prediction of genes (Prokka, PubMLST, ABRicate, ResFinder, Argannot, CARD, NCBI, VFDB, Snippy, CARD, PlasmidFinder, SeqSero)
Data output and visualization R-Studio
Core processes

Figure 1. Schematic representation of the workflow included in this study, the upstream process includes the preparation of the samples, sequencing and generation of the raw data while the core process summarised the bioinformatics analysis and the tools used for the data output and visualization of the results.

5. GENOMIC COMPARISON OF 22 ISOLATES

The genomes were identified using a hybrid approach, combining the results of the MLST typing scheme and the extraction of 16s gene sequencing. Details on the quality of the assemblies as determined with QUAST are presented in **Figure-2.**



		quality assessment																					
n isolate	Strain	# contigs (>= 0 bp)	# contigs (>= 1000 bp)	# contigs (>= 5000 bp)	# contigs (>= 10000 bp)	# contigs (>= 25000 bp)	# contigs (>= 50000 bp)	Total length (>= 0 bp)	Total length (>= 1000 bp)	Total length (>= 5000 bp)	Total length (>= 10000 bp)	Total length (>= 25000 bp)	Total length (>= 50000 bp)	# contigs	Largest contig	Total length	GC (%)	N50	N75	L50	L75	# N's per 100 kbp
1	CFS3535	48	35	27	26	23	18	4762069	4752937	4732373	4725307	4672738	4516097	48	661262	4762069	56.46	344176	161404	6	11	14.24
2	CFS3536	42	37	33	30	27	24	4592524	4588635	4580913	4564519	4521554	4427343	42	414782	4592524	56.86	267010	136001	7	13	4.25
3	CFS3537	52	35	29	27	24	19	4728652	4716618	4698919	4682245	4636062	4479530	52	682640	4728652	56.49	360593	150892	5	11	14.25
4	CFS3538	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
5	CFS3539	35	32	27	24	23	21	4506267	4504461	4495622	4477625	4458075	4393708	35	479484	4506267	56.95	299512	173069	6	11	4.24
6	CFS3540	44	39	32	29	27	24	4584737	4581423	4565987	4546432	4515043	4420732	44	414732	4584737	56.89	267010	135998	7	13	6.26
7	CFS3541	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
8	CFS3542	53	47	39	36	35	29	4591327	4587800	4566582	4544548	4524998	4320594	53	412389	4591327	56.87	188465	92932	9	17	0
9	CFS4391	37	29	21	21	20	19	5006977	5001291	4982135	4982135	4971907	4934369	37	1094456	5006977	51.95	255110	217726	6	11	7.87
10	CFS4392	39	31	22	22	20	17	5006270	5000504	4978748	4978748	4957483	4853401	39	1081186	5006270	51.95	323683	217726	5	10	9.79
11	CFS4393	37	29	21	21	19	18	5006913	5001147	4980554	4980554	4959289	4921751	37	1081329	5006913	51.95	285891	218104	6	11	7.85
12	CFS4394	47	35	29	29	27	24	5055415	5047363	5034033	5034033	5012978	4909020	47	869163	5055415	51.93	218104	162036	8	14	7.75
13	CFS4395	40	30	22	22	20	18	5005414	4998074	4978520	4978520	4957255	4869881	40	1081279	5005414	51.95	255086	217827	6	11	11.71
14	CFS4396	41	29	21	21	19	18	5005966	4997884	4978703	4978703	4957438	4919900	41	1081329	5005966	51.95	255110	217827	6	11	7.87
15	CFS4397	38	29	20	20	18	17	5005491	4999107	4978504	4978504	4957239	4919701	38	1081303	5005491	51.95	305166	217827	5	10	11.61
16	CFS4398	35	26	19	19	18	17	5008444	5001803	4983609	4983609	4973381	4935843	35	1081456	5008444	51.95	338990	228893	4	9	13.68
17	F2151	19	18	12	12	12	10	3051528	3050836	3040705	3040705	3040705	2960017	19	1254452	3051528	37.81	531720	127058	2	4	15.66
18	F2152	19	18	12	12	12	10	3052166	3051473	3041342	3041342	3041342	2960654	19	1254482	3052166	37.81	532224	127058	2	4	22.05
19	F2153	19	18	12	12	12	10	3051762	3051069	3040938	3040938	3040938	2960250	19	1254452	3051762	37.81	531950	127058	2	4	19.04
20	F2154	19	18	12	12	12	10	3051731	3051039	3040908	3040908	3040908	2960220	19	1254451	3051731	37.81	531924	127058	2	4	15.66
21	F2155	19	18	12	12	12	10	3051760	3051067	3040936	3040936	3040936	2960248	19	1254451	3051760	37.81	531949	127058	2	4	19.01
22	F2160	19	18	12	12	12	10	3052506	3051813	3041682	3041682	3041682	2960994	19	1254480	3052506	37.82	532666	127058	2	4	18.97
23	F2161	16	14	11	11	11	9	2953558	2952330	2946255	2946255	2946255	2870424	16	1254314	2953558	37.87	540972	464798	2	3	13.17
24	F2166	8	7	5	5	5	5	2919376	2918473	2913402	2913402	2913402	2913402	8	1563119	2919376	37.89	1563119	503407	1	3	16.41

Figure-2. Quality assessment of the 24 isolates included in this study. The figure includes the values of the size of the contigs, GC contents (%) and parameters on the quality of the sequences (N75, L50 and L75). Two samples are not included (n.a: not appliable).

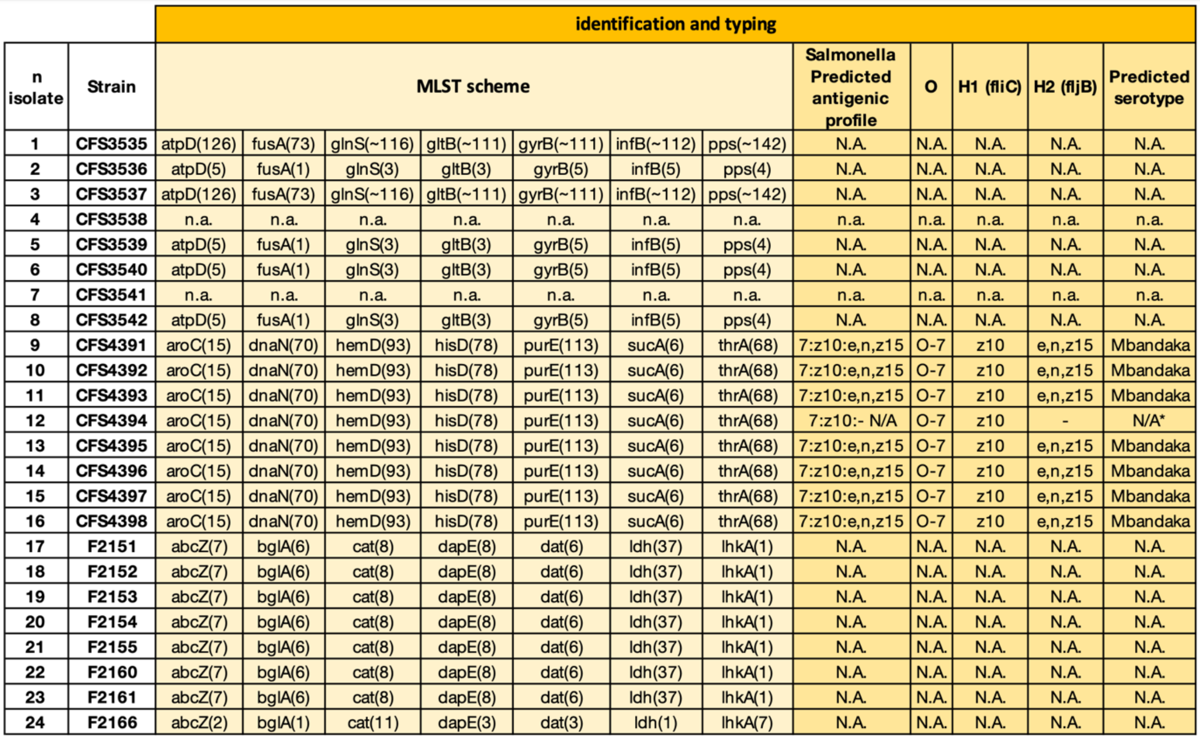
Two samples were identified as *Franconibacter sp.,* which are bacteria genetically related to *Cronobacter sakazakii* (identified in four samples). Eight samples resulted *Salmonella enterica* and eight isolates were identified as *Listeria monocytogenes* **(Table 2).**

Table 2. Details on the identification of the isolates included in the study and MLST results

n isolate	Strain	Species identified	MLST	n isolate	Strain	Species identified	MLST	n isolate	Strain	Species identified	MLST
1	**CFS3535**	*Franconibacter sp.*	n.d.	**9**	**CFS4391**	*Salmonella enterica*	413	**17**	**F2151**	*Listeria monocytogenes*	121
2	**CFS3536**	*Cronobacter sakazakii*	4	**10**	**CFS4392**	*Salmonella enterica*	413	**18**	**F2152**	*Listeria monocytogenes*	121
3	**CFS3537**	*Franconibacter sp.*	n.d.	**11**	**CFS4393**	*Salmonella enterica*	413	**19**	**F2153**	*Listeria monocytogenes*	121
4	**CFS3538**	*not included*	n.a.	**12**	**CFS4394**	*Salmonella enterica*	413	**20**	**F2154**	*Listeria monocytogenes*	121
5	**CFS3539**	*Cronobacter sakazakii*	4	**13**	**CFS4395**	*Salmonella enterica*	413	**21**	**F2155**	*Listeria monocytogenes*	121
6	**CFS3540**	*Cronobacter sakazakii*	4	**14**	**CFS4396**	*Salmonella enterica*	413	**22**	**F2160**	*Listeria monocytogenes*	121
7	**CFS3541**	*not included*	n.a.	**15**	**CFS4397**	*Salmonella enterica*	413	**23**	**F2161**	*Listeria monocytogenes*	121
8	**CFS3542**	*Cronobacter sakazakii*	4	**16**	**CFS4398**	*Salmonella enterica*	413	**24**	**F2166**	*Listeria monocytogenes*	5

n.d.: not determined (the typing scheme could not assign a ST)
n.a.: not appliable, the strain was not analysed because not all the laboratories were able to sequence the isolate (not included in the analysis).

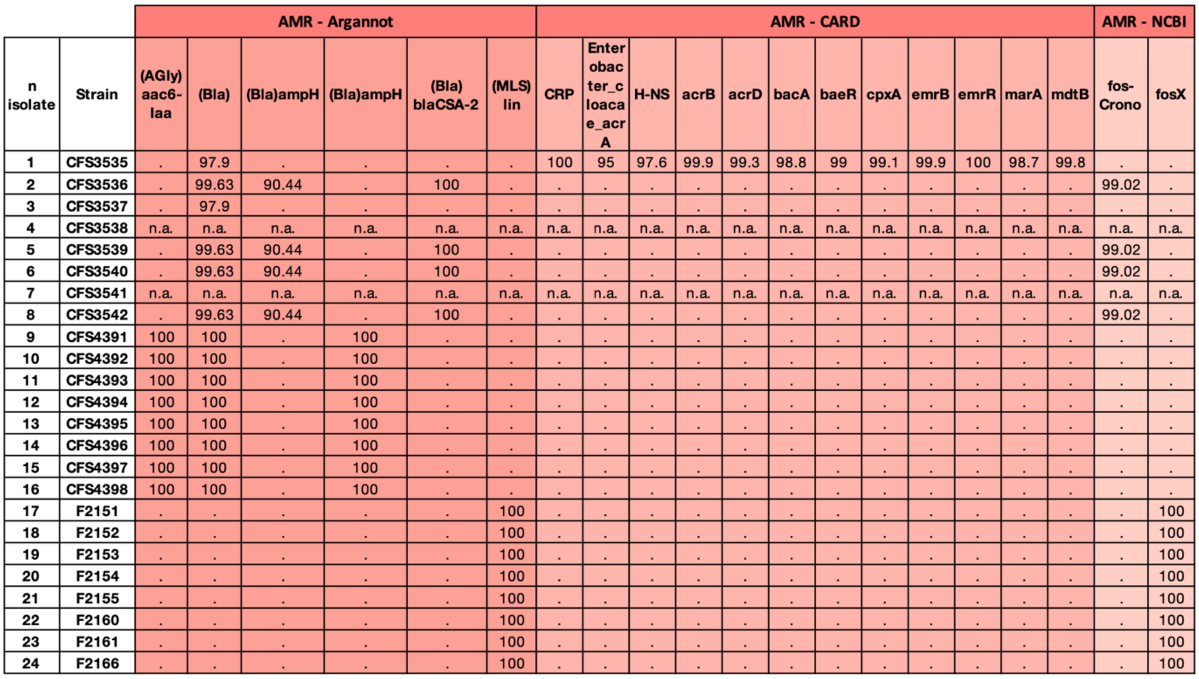
The genes identified for the identification of the ST are presented in **Figure-3** which include the typing scheme for salmonella and the prediction of the serotype.



		identification and typing											
n isolate	Strain	MLST scheme							Salmonella Predicted antigenic profile	O	H1 (fliC)	H2 (fljB)	Predicted serotype
1	CFS3535	atpD(126)	fusA(73)	glnS(~116)	gltB(~111)	gyrB(~111)	infB(~112)	pps(~142)	N.A.	N.A.	N.A.	N.A.	N.A.
2	CFS3536	atpD(5)	fusA(1)	glnS(3)	gltB(3)	gyrB(5)	infB(5)	pps(4)	N.A.	N.A.	N.A.	N.A.	N.A.
3	CFS3537	atpD(126)	fusA(73)	glnS(~116)	gltB(~111)	gyrB(~111)	infB(~112)	pps(~142)	N.A.	N.A.	N.A.	N.A.	N.A.
4	CFS3538	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
5	CFS3539	atpD(5)	fusA(1)	glnS(3)	gltB(3)	gyrB(5)	infB(5)	pps(4)	N.A.	N.A.	N.A.	N.A.	N.A.
6	CFS3540	atpD(5)	fusA(1)	glnS(3)	gltB(3)	gyrB(5)	infB(5)	pps(4)	N.A.	N.A.	N.A.	N.A.	N.A.
7	CFS3541	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
8	CFS3542	atpD(5)	fusA(1)	glnS(3)	gltB(3)	gyrB(5)	infB(5)	pps(4)	N.A.	N.A.	N.A.	N.A.	N.A.
9	CFS4391	aroC(15)	dnaN(70)	hemD(93)	hisD(78)	purE(113)	sucA(6)	thrA(68)	7:z10:e,n,z15	O-7	z10	e,n,z15	Mbandaka
10	CFS4392	aroC(15)	dnaN(70)	hemD(93)	hisD(78)	purE(113)	sucA(6)	thrA(68)	7:z10:e,n,z15	O-7	z10	e,n,z15	Mbandaka
11	CFS4393	aroC(15)	dnaN(70)	hemD(93)	hisD(78)	purE(113)	sucA(6)	thrA(68)	7:z10:e,n,z15	O-7	z10	e,n,z15	Mbandaka
12	CFS4394	aroC(15)	dnaN(70)	hemD(93)	hisD(78)	purE(113)	sucA(6)	thrA(68)	7:z10:- N/A	O-7	z10	-	N/A*
13	CFS4395	aroC(15)	dnaN(70)	hemD(93)	hisD(78)	purE(113)	sucA(6)	thrA(68)	7:z10:e,n,z15	O-7	z10	e,n,z15	Mbandaka
14	CFS4396	aroC(15)	dnaN(70)	hemD(93)	hisD(78)	purE(113)	sucA(6)	thrA(68)	7:z10:e,n,z15	O-7	z10	e,n,z15	Mbandaka
15	CFS4397	aroC(15)	dnaN(70)	hemD(93)	hisD(78)	purE(113)	sucA(6)	thrA(68)	7:z10:e,n,z15	O-7	z10	e,n,z15	Mbandaka
16	CFS4398	aroC(15)	dnaN(70)	hemD(93)	hisD(78)	purE(113)	sucA(6)	thrA(68)	7:z10:e,n,z15	O-7	z10	e,n,z15	Mbandaka
17	F2151	abcZ(7)	bglA(6)	cat(8)	dapE(8)	dat(6)	ldh(37)	lhkA(1)	N.A.	N.A.	N.A.	N.A.	N.A.
18	F2152	abcZ(7)	bglA(6)	cat(8)	dapE(8)	dat(6)	ldh(37)	lhkA(1)	N.A.	N.A.	N.A.	N.A.	N.A.
19	F2153	abcZ(7)	bglA(6)	cat(8)	dapE(8)	dat(6)	ldh(37)	lhkA(1)	N.A.	N.A.	N.A.	N.A.	N.A.
20	F2154	abcZ(7)	bglA(6)	cat(8)	dapE(8)	dat(6)	ldh(37)	lhkA(1)	N.A.	N.A.	N.A.	N.A.	N.A.
21	F2155	abcZ(7)	bglA(6)	cat(8)	dapE(8)	dat(6)	ldh(37)	lhkA(1)	N.A.	N.A.	N.A.	N.A.	N.A.
22	F2160	abcZ(7)	bglA(6)	cat(8)	dapE(8)	dat(6)	ldh(37)	lhkA(1)	N.A.	N.A.	N.A.	N.A.	N.A.
23	F2161	abcZ(7)	bglA(6)	cat(8)	dapE(8)	dat(6)	ldh(37)	lhkA(1)	N.A.	N.A.	N.A.	N.A.	N.A.
24	F2166	abcZ(2)	bglA(1)	cat(11)	dapE(3)	dat(3)	ldh(1)	lhkA(7)	N.A.	N.A.	N.A.	N.A.	N.A.

Figure-3. Genes identified for the detection of the sequence types according to the MLST scheme and for ins-silico the serotyping of *Salmonella*. Two samples are not included (n.a: not appliable). N/A* The predicted antigenic profile does not exist in the White-Kauffmann-Le Minor scheme; N.A. not applicable.

The assembled sequences were analysed for the presence of AMR genes using different databases (Argannot, CARD, NCBI and ResFinder). The unique results are presented in **Figure-4.**



		AMR - Argannot						AMR - CARD												AMR - NCBI	
n isolate	Strain	(AGly) aac6-Iaa	(Bla)	(Bla)ampH	(Bla)ampH	(Bla) blaCSA-2	(MLS) lin	CRP	Enterobacter_cloacae_acrA	H-NS	acrB	acrD	bacA	baeR	cpxA	emrB	emrR	marA	mdtB	fos-Crono	fosX
1	CFS3535	.	97.9	.	.	.	.	100	95	97.6	99.9	99.3	98.8	99	99.1	99.9	100	98.7	99.8	.	.
2	CFS3536	.	99.63	90.44	.	100	.	.	.	.	.	.	.	.	.	.	.	.	.	99.02	.
3	CFS3537	.	97.9	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
4	CFS3538	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
5	CFS3539	.	99.63	90.44	.	100	.	.	.	.	.	.	.	.	.	.	.	.	.	99.02	.
6	CFS3540	.	99.63	90.44	.	100	.	.	.	.	.	.	.	.	.	.	.	.	.	99.02	.
7	CFS3541	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
8	CFS3542	.	99.63	90.44	.	100	.	.	.	.	.	.	.	.	.	.	.	.	.	99.02	.
9	CFS4391	100	100	.	100	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
10	CFS4392	100	100	.	100	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
11	CFS4393	100	100	.	100	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
12	CFS4394	100	100	.	100	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
13	CFS4395	100	100	.	100	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
14	CFS4396	100	100	.	100	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
15	CFS4397	100	100	.	100	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
16	CFS4398	100	100	.	100	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
17	F2151	.	.	.	.	.	100	.	.	.	.	.	.	.	.	.	.	.	.	.	100
18	F2152	.	.	.	.	.	100	.	.	.	.	.	.	.	.	.	.	.	.	.	100
19	F2153	.	.	.	.	.	100	.	.	.	.	.	.	.	.	.	.	.	.	.	100
20	F2154	.	.	.	.	.	100	.	.	.	.	.	.	.	.	.	.	.	.	.	100
21	F2155	.	.	.	.	.	100	.	.	.	.	.	.	.	.	.	.	.	.	.	100
22	F2160	.	.	.	.	.	100	.	.	.	.	.	.	.	.	.	.	.	.	.	100
23	F2161	.	.	.	.	.	100	.	.	.	.	.	.	.	.	.	.	.	.	.	100
24	F2166	.	.	.	.	.	100	.	.	.	.	.	.	.	.	.	.	.	.	.	100

Figure-4. AMR genes identified among the isolates. Different databases were used and the unique results are presented in the figure. The value correspond to the percentage of coverage to the reference sequences. Two samples are not included (n.a: not appliable).

The presence of plasmids was predicted with the use of the tool PlasmidFinder and were detected multiple marker genes on all the *Salmonella* spp. isolates, that carrying AMR-resistance genes, including RepA_1_pKPC-CAV1321, IncHI2A_1 and IncHI2_1. The marker IncL/M(pMU407)_1_pMU407 was also identified in the isolates 12 (CFS4394). All the *L. monocytogenes* resulted negative for the presence of plasmids while the ColRNAI_1 plasmid was detected in the assemblies of both *C. sakazakii and Franconibacter spp* while Col440II_1 and IncFIB(pCTU1)_1_pCTU1 were detected only in the four *C. sakazakii* isolates **(Table 3).**

Table 3. Plasmid identified among the 22 isolates.

Plasmid	Isolates
Col440II_1	CFS3536, CFS3539, CFS3540, CFS3542
ColRNAI_1	CFS3535, CFS3536, CFS3537, CFS3539, CFS3540, CFS3542
IncFIB(pCTU1)_1_pCTU1	CFS3536, CFS3539, CFS3540, CFS3542
IncHI2A_1	CFS4391, CFS4392, CFS4393, CFS4394, CFS4395, CFS4396, CFS4397, CFS4398
IncHI2_1	CFS4391, CFS4392, CFS4393, CFS4394, CFS4395, CFS4396, CFS4397, CFS4398
IncL/M(pMU407)_1_pMU407	CFS4394
RepA_1_pKPC-CAV1321	CFS4391, CFS4392, CFS4393, CFS4394, CFS4395, CFS4396, CFS4397, CFS4398

In order to represent the genomic relationship of the isolates, the samples were grouped according to the identified species, with genomes falling into multiple groups of only remotely related genomes (Figure-5): group 1, *Listeria monocytogenes;* group 2*, Salmonella enterica;* and group 3 *Franconibacter* sp. And *Cronobacter sakazakii.* The group of *Salmonella* spp. was identified as *Salmonella enterica* and it was possible to assign the serotype Mbandaka according to the prediction of the White-Kauffmann-Le Minor scheme for seven out of eight isolates. One sample (CFS4394) resulted as a non-existent profile.

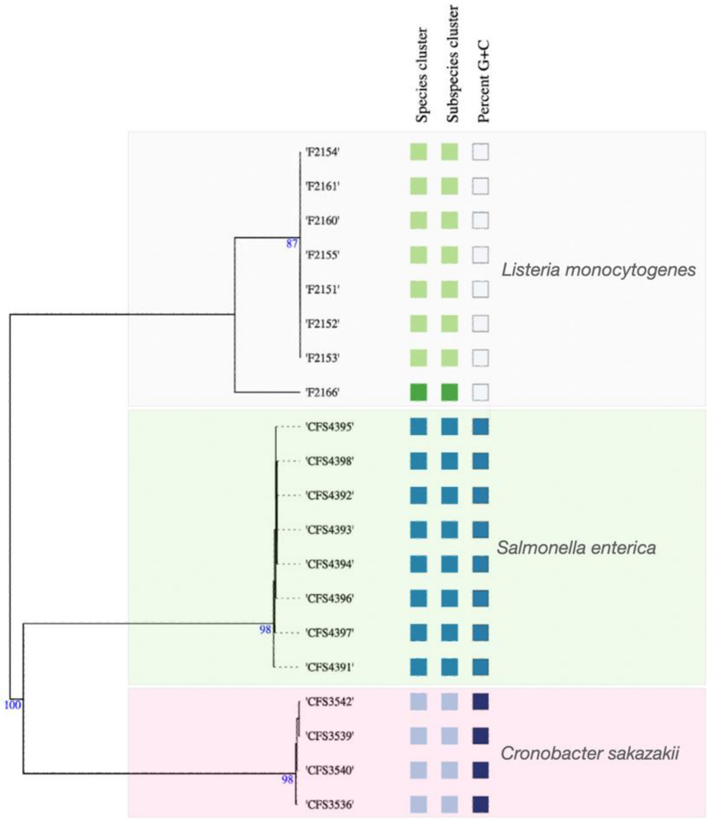


Figure-5. Phylogenetic tree representing 20 genomes (CFS-indexed) of the three major species included in the study.

In group 3, two *Franconibacter* spp*.* were identifed and it was not possible to assign the MLST. *C. sakazakii* were identified as ST4 **(Table 2)*.*** The samples included in group 3 were represented in the phylogram (phylogenetic tree) **Figure-6.**

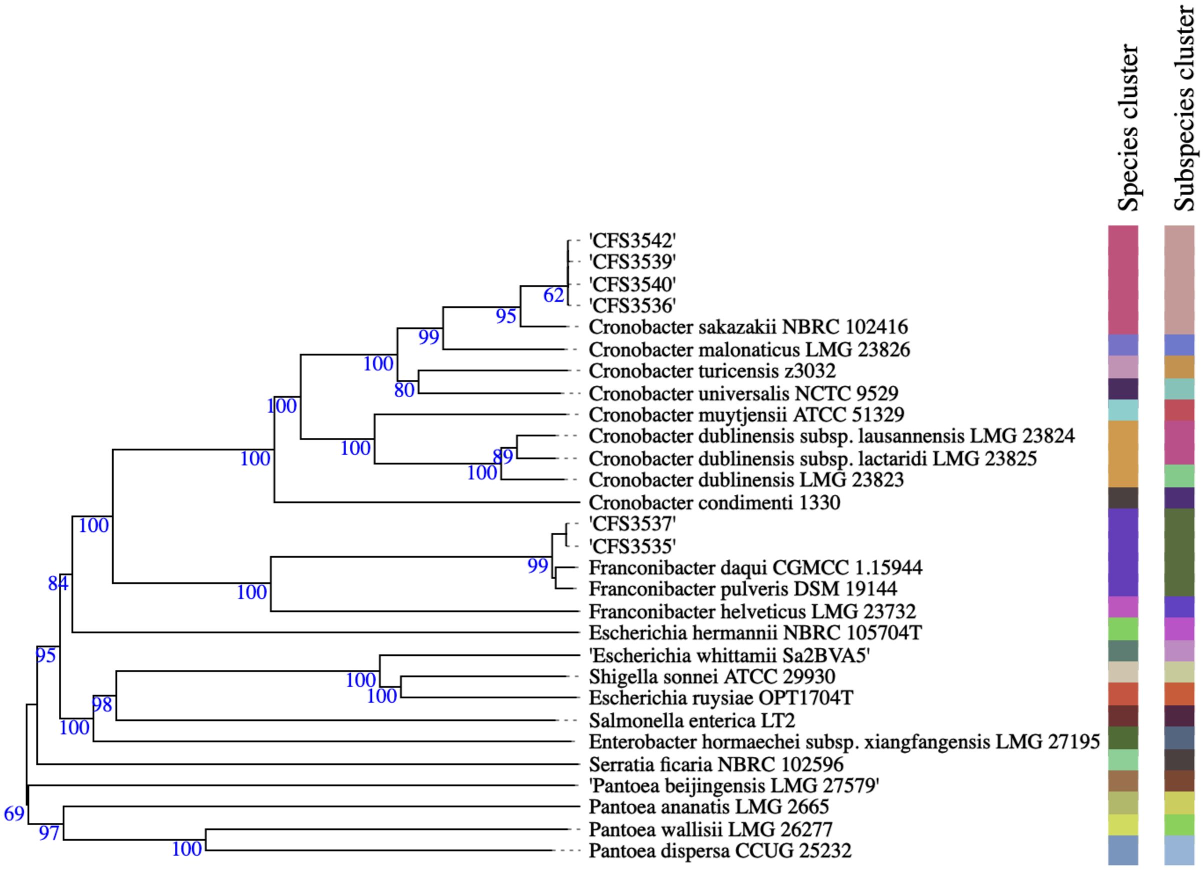

Species cluster
Subspecies cluster
'CFS3542'
'CFS3539'
'CFS3540'
'CFS3536'
Cronobacter sakazakii NBRC 102416
Cronobacter malonaticus LMG 23826
Cronobacter turicensis z3032
Cronobacter universalis NCTC 9529
Cronobacter muytjensii ATCC 51329
Cronobacter dublinensis subsp. lausannensis LMG 23824
Cronobacter dublinensis subsp. lactaridi LMG 23825
Cronobacter dublinensis LMG 23823
Cronobacter condimenti 1330
'CFS3537'
'CFS3535'
Franconibacter daqui CGMCC 1.15944
Franconibacter pulveris DSM 19144
Franconibacter helveticus LMG 23732
Escherichia hermannii NBRC 105704T
'Escherichia whittamii Sa2BVA5'
Shigella sonnei ATCC 29930
Escherichia ruysiae OPT1704T
Salmonella enterica LT2
Enterobacter hormaechei subsp. xiangfangensis LMG 27195
Serratia ficaria NBRC 102596
'Pantoea beijingensis LMG 27579'
Pantoea ananatis LMG 2665
Pantoea wallisii LMG 26277
Pantoea dispersa CCUG 25232
62
95
99
100
80
100
100
100
89
100
100
100
99
100
84
95
100
100
98
100
69
97
100

Figure-6. Phylogenetic tree representing the 6 genomes (CFS-indexed) included in the study and reference genome gnomically related.

6. GENOMIC COMPARISON BETWEEN LABORATORIES

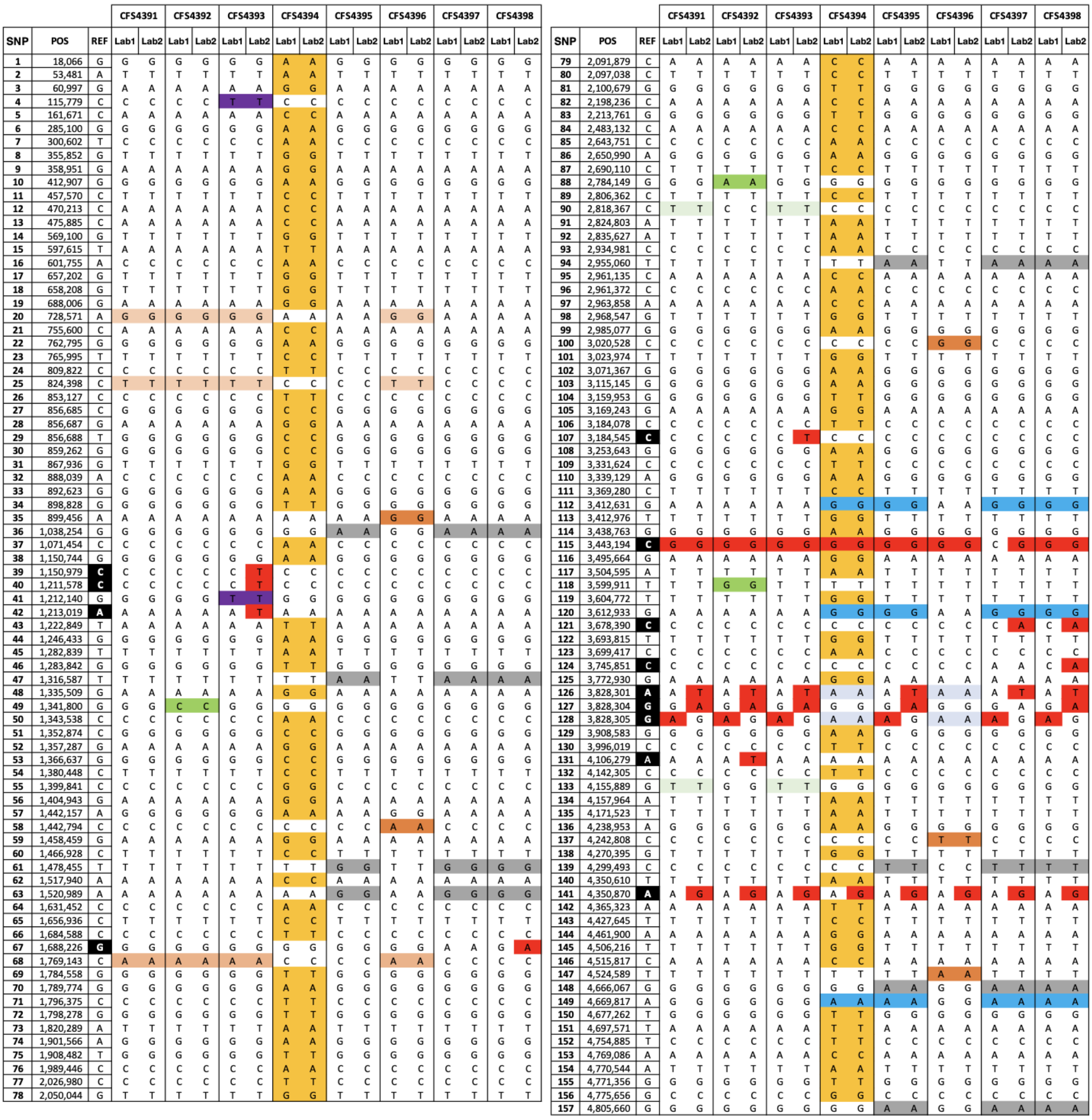
The analysis was carried out in two different laboratories. Results from both laboratories corresponded and all genomic markers were assigned correctly.

In order to further-analyse the possible differences or biases resulting from technical issues, the analysis of the SNPs was implemented in the 22 samples. Considering the complexity and the genomic differences among the three groups of bacteria, the *Salmonella* spp. group was chosen for the lower diversity of the genomes. Snippy was used for finding SNPs between the reference genome (NC_003197.2 - *Salmonella enterica* subsp. *enterica* serovar Typhimurium str. LT2) and the data of three different levels:

- *Raw reads* – non-manipulated data, generated by sequencing machines (fastq format);
- *Cleaned reads* – data filtered out for low quality and short reads, low quality bases, removal of adapter sequences (fastq format);
- *Assembl*ed genomes – draft genomes assembled using the cleaned reads as input. The data were processed with SPAdes and contigs smaller than 500 bases were filtered out (fasta format).

The software used for the SNPs detection, Snippy (Seemann, 2018), was able to find both substitutions (snps) and insertions/deletions (indels). Sets of Snippy results were used to generate core SNP alignments and ultimately phylogenomic trees. In addition, the algorithm Gubbins (Genealogies Unbiased By recomBinations In Nucleotide Sequences) was used for constructing a phylogeny based on the putative point mutations outside of regions containing elevated densities of base substitutions (Croucher et al., 2015) in the case of the collection of microorganisms used for this study.

The three approaches provided matching results, in fact a total of 45,950 substitutions were found when comparing the 16 samples with the reference strain. However, the substitutions found among the samples amounted to 157 SNPs identified among the 8 isolates during the exercise between the two laboratories (Lab1 and Lab2), (Figure-7) across the three approaches (raw reads, cleaned reads and assembled genomes). The patterns of SNPs are represented with different colours where SNPs are matching in the two laboratories, while the SNPs coloured in red are identified only in one laboratory (13 cases). Interestingly, these cases were detected among the three different bioinformatics approaches, suggesting a possible mutation of the sub-cultured strains in the two laboratories. 11 cases out of the 13 were identified in the laboratory 2 which did not perform the laboratory analysis during the same time frame as laboratory 1.



			CFS4391		CFS4392		CFS4393		CFS4394		CFS4395		CFS4396		CFS4397		CFS4398	
SNP	POS	REF	Lab1	Lab2	Lab1	Lab2	Lab1	Lab2	Lab1	Lab2	Lab1	Lab2	Lab1	Lab2	Lab1	Lab2	Lab1	Lab2
1	18,066	G	G	G	G	G	G	G	A	A	G	G	G	G	G	G	G	G
2	53,481	A	T	T	T	T	T	T	A	A	T	T	T	T	T	T	T	T
3	60,997	G	A	A	A	A	A	A	G	G	A	A	A	A	A	A	A	A
4	115,779	C	C	C	C	C	T	T	C	C	C	C	C	C	C	C	C	C
5	161,671	C	A	A	A	A	A	A	C	C	A	A	A	A	A	A	A	A
6	285,100	G	G	G	G	G	G	G	A	A	G	G	G	G	G	G	G	G
7	300,602	T	C	C	C	C	C	C	A	A	C	C	C	C	C	C	C	C
8	355,852	G	T	T	T	T	T	T	G	G	T	T	T	T	T	T	T	T
9	358,951	G	A	A	A	A	A	A	G	G	A	A	A	A	A	A	A	A
10	412,907	G	G	G	G	G	G	G	A	A	G	G	G	G	G	G	G	G
11	457,570	C	T	T	T	T	T	T	C	C	T	T	T	T	T	T	T	T
12	470,213	C	A	A	A	A	A	A	C	C	A	A	A	A	A	A	A	A
13	475,885	C	A	A	A	A	A	A	C	C	A	A	A	A	A	A	A	A
14	569,100	G	T	T	T	T	T	T	G	G	T	T	T	T	T	T	T	T
15	597,615	T	A	A	A	A	A	A	T	T	A	A	A	A	A	A	A	A
16	601,755	A	C	C	C	C	C	C	A	A	C	C	C	C	C	C	C	C
17	657,202	G	T	T	T	T	T	T	G	G	T	T	T	T	T	T	T	T
18	658,208	G	T	T	T	T	T	T	G	G	T	T	T	T	T	T	T	T
19	688,006	G	A	A	A	A	A	A	G	G	A	A	A	A	A	A	A	A
20	728,571	A	G	G	G	G	G	G	A	A	A	A	G	G	A	A	A	A
21	755,600	C	A	A	A	A	A	A	C	C	A	A	A	A	A	A	A	A
22	762,795	G	G	G	G	G	G	G	A	A	G	G	G	G	G	G	G	G
23	765,995	T	T	T	T	T	T	T	C	C	T	T	T	T	T	T	T	T
24	809,822	C	C	C	C	C	C	C	T	T	C	C	C	C	C	C	C	C
25	824,398	C	T	T	T	T	T	T	C	C	C	C	T	T	C	C	C	C
26	853,127	C	C	C	C	C	C	C	T	T	C	C	C	C	C	C	C	C
27	856,685	C	G	G	G	G	G	G	C	C	G	G	G	G	G	G	G	G
28	856,687	G	A	A	A	A	A	A	G	G	A	A	A	A	A	A	A	A
29	856,688	T	G	G	G	G	G	G	C	C	G	G	G	G	G	G	G	G
30	859,262	G	G	G	G	G	G	G	C	C	G	G	G	G	G	G	G	G
31	867,936	G	T	T	T	T	T	T	G	G	T	T	T	T	T	T	T	T
32	888,039	A	C	C	C	C	C	C	A	A	C	C	C	C	C	C	C	C
33	892,623	G	G	G	G	G	G	G	A	A	G	G	G	G	G	G	G	G
34	898,828	G	G	G	G	G	G	G	T	T	G	G	G	G	G	G	G	G
35	899,456	A	A	A	A	A	A	A	A	A	A	A	G	G	A	A	A	A
36	1,038,254	G	G	G	G	G	G	G	G	G	A	A	G	G	A	A	A	A
37	1,071,454	C	C	C	C	C	C	C	A	A	C	C	C	C	C	C	C	C
38	1,150,744	G	G	G	G	G	G	G	A	A	G	G	G	G	G	G	G	G
39	1,150,979	C	C	C	C	C	C	T	C	C	C	C	C	C	C	C	C	C
40	1,211,578	C	C	C	C	C	C	T	C	C	C	C	C	C	C	C	C	C
41	1,212,140	G	G	G	G	G	T	T	G	G	G	G	G	G	G	G	G	G
42	1,213,019	A	A	A	A	A	A	T	A	A	A	A	A	A	A	A	A	A
43	1,222,849	T	A	A	A	A	A	A	T	T	A	A	A	A	A	A	A	A
44	1,246,433	G	G	G	G	G	G	G	A	A	G	G	G	G	G	G	G	G
45	1,282,839	T	T	T	T	T	T	T	A	A	T	T	T	T	T	T	T	T
46	1,283,842	G	G	G	G	G	G	G	T	T	G	G	G	G	G	G	G	G
47	1,316,587	T	T	T	T	T	T	T	T	T	A	A	T	T	A	A	A	A
48	1,335,509	G	A	A	A	A	A	A	G	G	A	A	A	A	A	A	A	A
49	1,341,800	G	G	G	C	C	G	G	G	G	G	G	G	G	G	G	G	G
50	1,343,538	C	C	C	C	C	C	C	A	A	C	C	C	C	C	C	C	C
51	1,352,874	C	G	G	G	G	G	G	C	C	G	G	G	G	G	G	G	G
52	1,357,287	G	A	A	A	A	A	A	G	G	A	A	A	A	A	A	A	A
53	1,366,637	G	G	G	G	G	G	G	C	C	G	G	G	G	G	G	G	G
54	1,380,448	C	T	T	T	T	T	T	C	C	T	T	T	T	T	T	T	T
55	1,399,841	C	C	C	C	C	C	C	G	G	C	C	C	C	C	C	C	C
56	1,404,943	G	A	A	A	A	A	A	G	G	A	A	A	A	A	A	A	A
57	1,442,157	A	G	G	G	G	G	G	A	A	A	A	G	G	A	A	A	A
58	1,442,794	C	C	C	C	C	C	C	C	C	C	C	A	A	C	C	C	C
59	1,458,459	G	A	A	A	A	A	A	G	G	A	A	A	A	A	A	A	A
60	1,466,928	C	T	T	T	T	T	T	C	C	T	T	T	T	T	T	T	T
61	1,478,455	T	T	T	T	T	T	T	T	T	G	G	T	T	G	G	G	G
62	1,517,940	A	A	A	A	A	A	A	C	C	A	A	A	A	A	A	A	A
63	1,520,989	A	A	A	A	A	A	A	A	A	G	G	A	A	G	G	G	G
64	1,631,452	C	C	C	C	C	C	C	A	A	C	C	C	C	C	C	C	C
65	1,656,936	C	T	T	T	T	T	T	C	C	T	T	T	T	T	T	T	T
66	1,684,588	C	C	C	C	C	C	C	T	T	C	C	C	C	C	C	C	C
67	1,688,226	G	G	G	G	G	G	G	G	G	G	G	G	G	A	A	G	A
68	1,769,143	C	A	A	A	A	A	A	C	C	C	C	A	A	C	C	C	C
69	1,784,558	G	G	G	G	G	G	G	T	T	G	G	G	G	G	G	G	G
70	1,789,774	G	G	G	G	G	G	G	A	A	G	G	G	G	G	G	G	G
71	1,796,375	C	C	C	C	C	C	C	T	T	C	C	C	C	C	C	C	C
72	1,798,278	G	G	G	G	G	G	G	T	T	G	G	G	G	G	G	G	G
73	1,820,289	A	T	T	T	T	T	T	A	A	T	T	T	T	T	T	T	T
74	1,901,566	A	G	G	G	G	G	G	A	A	G	G	G	G	G	G	G	G
75	1,908,482	T	G	G	G	G	G	G	T	T	G	G	G	G	G	G	G	G
76	1,989,446	C	C	C	C	C	C	C	A	A	C	C	C	C	C	C	C	C
77	2,026,980	C	C	C	C	C	C	C	T	T	C	C	C	C	C	C	C	C
78	2,050,044	G	T	T	T	T	T	T	G	G	T	T	T	T	T	T	T	T
79	2,091,879	C	A	A	A	A	A	A	C	C	A	A	A	A	A	A	A	A
80	2,097,038	C	T	T	T	T	T	T	C	C	T	T	T	T	T	T	T	T
81	2,100,679	G	G	G	G	G	G	G	T	T	G	G	G	G	G	G	G	G
82	2,198,236	C	A	A	A	A	A	A	C	C	A	A	A	A	A	A	A	A
83	2,213,761	G	G	G	G	G	G	G	T	T	G	G	G	G	G	G	G	G
84	2,483,132	C	A	A	A	A	A	A	C	C	A	A	A	A	A	A	A	A
85	2,643,751	C	C	C	C	C	C	C	A	A	C	C	C	C	C	C	C	C
86	2,650,990	A	G	G	G	G	G	G	A	A	G	G	G	G	G	G	G	G
87	2,690,110	C	T	T	T	T	T	T	C	C	T	T	T	T	T	T	T	T
88	2,784,149	G	G	G	A	A	G	G	G	G	G	G	G	G	G	G	G	G
89	2,806,362	C	T	T	T	T	T	T	C	C	T	T	T	T	T	T	T	T
90	2,818,367	C	T	T	C	C	T	T	C	C	C	C	C	C	C	C	C	C
91	2,824,803	A	T	T	T	T	T	T	A	A	T	T	T	T	T	T	T	T
92	2,835,627	A	T	T	T	T	T	T	A	A	T	T	T	T	T	T	T	T
93	2,934,981	C	C	C	C	C	C	C	A	A	C	C	C	C	C	C	C	C
94	2,955,060	T	T	T	T	T	T	T	T	T	A	A	T	T	A	A	A	A
95	2,961,135	C	A	A	A	A	A	A	C	C	A	A	A	A	A	A	A	A
96	2,961,372	C	C	C	C	C	C	C	A	A	C	C	C	C	C	C	C	C
97	2,963,858	A	A	A	A	A	A	A	C	C	A	A	A	A	A	A	A	A
98	2,968,547	G	T	T	T	T	T	T	G	G	T	T	T	T	T	T	T	T
99	2,985,077	G	G	G	G	G	G	G	A	A	G	G	G	G	G	G	G	G
100	3,020,528	C	C	C	C	C	C	C	C	C	C	C	G	G	C	C	C	C
101	3,023,974	T	T	T	T	T	T	T	G	G	T	T	T	T	T	T	T	T
102	3,071,367	G	G	G	G	G	G	G	A	A	G	G	G	G	G	G	G	G
103	3,115,145	G	G	G	G	G	G	G	A	A	G	G	G	G	G	G	G	G
104	3,159,953	G	G	G	G	G	G	G	T	T	G	G	G	G	G	G	G	G
105	3,169,243	G	A	A	A	A	A	A	G	G	A	A	A	A	A	A	A	A
106	3,184,078	C	C	C	C	C	C	C	T	T	C	C	C	C	C	C	C	C
107	3,184,545	C	C	C	C	C	C	T	C	C	C	C	C	C	C	C	C	C
108	3,253,643	G	G	G	G	G	G	G	A	A	G	G	G	G	G	G	G	G
109	3,331,624	C	C	C	C	C	C	C	T	T	C	C	C	C	C	C	C	C
110	3,339,129	A	G	G	G	G	G	G	A	A	G	G	G	G	G	G	G	G
111	3,369,280	C	T	T	T	T	T	T	C	C	T	T	T	T	T	T	T	T
112	3,412,631	G	A	A	A	A	A	A	G	G	G	G	A	A	G	G	G	G
113	3,412,976	T	T	T	T	T	T	T	G	G	T	T	T	T	T	T	T	T
114	3,438,763	G	G	G	G	G	G	G	A	A	G	G	G	G	G	G	G	G
115	3,443,194	C	G	G	G	G	G	G	G	G	G	G	G	G	C	G	G	G
116	3,495,664	G	A	A	A	A	A	A	G	G	A	A	A	A	A	A	A	A
117	3,504,595	A	T	T	T	T	T	T	A	A	T	T	T	T	T	T	T	T
118	3,599,911	T	T	T	G	G	T	T	T	T	T	T	T	T	T	T	T	T
119	3,604,772	T	T	T	T	T	T	T	G	G	T	T	T	T	T	T	T	T
120	3,612,933	G	A	A	A	A	A	A	G	G	G	G	A	A	G	G	G	G
121	3,678,390	C	C	C	C	C	C	C	C	C	C	C	C	C	C	A	C	A
122	3,693,815	T	T	T	T	T	T	T	G	G	T	T	T	T	T	T	T	T
123	3,699,417	C	C	C	C	C	C	C	A	A	C	C	C	C	C	C	C	C
124	3,745,851	C	C	C	C	C	C	C	C	C	C	C	C	C	A	A	C	A
125	3,772,930	G	A	A	A	A	A	A	G	G	A	A	A	A	A	A	A	A
126	3,828,301	A	A	T	A	T	A	T	A	A	A	T	A	A	A	T	A	T
127	3,828,304	G	G	A	G	A	G	A	G	G	G	A	G	G	G	A	G	A
128	3,828,305	G	A	G	A	G	A	G	A	A	A	G	A	A	A	G	A	G
129	3,908,583	G	G	G	G	G	G	G	A	A	G	G	G	G	G	G	G	G
130	3,996,019	C	C	C	C	C	C	C	T	T	C	C	C	C	C	C	C	C
131	4,106,279	A	A	A	A	T	A	A	A	A	A	A	A	A	A	A	A	A
132	4,142,305	C	C	C	C	C	C	C	T	T	C	C	C	C	C	C	C	C
133	4,155,889	G	T	T	G	G	T	T	G	G	G	G	G	G	G	G	G	G
134	4,157,964	A	T	T	T	T	T	T	A	A	T	T	T	T	T	T	T	T
135	4,171,523	T	T	T	T	T	T	T	A	A	T	T	T	T	T	T	T	T
136	4,238,953	A	G	G	G	G	G	G	A	A	G	G	G	G	G	G	G	G
137	4,242,808	C	C	C	C	C	C	C	C	C	C	C	T	T	C	C	C	C
138	4,270,395	G	T	T	T	T	T	T	G	G	T	T	T	T	T	T	T	T
139	4,299,493	C	C	C	C	C	C	C	C	C	T	T	C	C	T	T	T	T
140	4,350,610	T	T	T	T	T	T	T	A	A	T	T	T	T	T	T	T	T
141	4,350,870	A	A	G	A	G	A	G	A	G	A	G	A	G	A	G	A	G
142	4,365,323	A	A	A	A	A	A	A	T	T	A	A	A	A	A	A	A	A
143	4,427,645	T	T	T	T	T	T	T	C	C	T	T	T	T	T	T	T	T
144	4,461,900	A	A	A	A	A	A	A	G	G	A	A	A	A	A	A	A	A
145	4,506,216	T	T	T	T	T	T	T	G	G	T	T	T	T	T	T	T	T
146	4,515,817	C	A	A	A	A	A	A	C	C	A	A	A	A	A	A	A	A
147	4,524,589	T	T	T	T	T	T	T	T	T	T	T	A	A	T	T	T	T
148	4,666,067	G	G	G	G	G	G	G	G	G	A	A	G	G	A	A	A	A
149	4,669,817	A	G	G	G	G	G	G	A	A	A	A	G	G	A	A	A	A
150	4,677,262	T	G	G	G	G	G	G	T	T	G	G	G	G	G	G	G	G
151	4,697,571	T	A	A	A	A	A	A	T	T	A	A	A	A	A	A	A	A
152	4,754,885	T	C	C	C	C	C	C	T	T	C	C	C	C	C	C	C	C
153	4,769,086	A	A	A	A	A	A	A	C	C	A	A	A	A	A	A	A	A
154	4,770,544	A	T	T	T	T	T	T	A	A	T	T	T	T	T	T	T	T
155	4,771,356	G	G	G	G	G	G	G	T	T	G	G	G	G	G	G	G	G
156	4,775,656	G	C	C	C	C	C	C	G	G	C	C	C	C	C	C	C	C
157	4,805,660	G	G	G	G	G	G	G	G	G	A	A	G	G	A	A	A	A

Figure-7. SNPs identified among the 8 isolates during the exercise between the two laboratories (Lab1 and Lab2). The patterns of SNPs are represented with different colours where SNPs are matching in the two laboratories, while the SNPs coloured in red are identified only in one laboratory. Those highlighted in black indicate those corresponding to the original base substituted (SNP) in one laboratory only. In the figure, the position of the SNP (POS) and the corresponding base substituted from the reference (REF), strain NC_003197.2 - *Salmonella enterica* subsp. *enterica* serovar Typhimurium str. LT2, is indicated.

7. CONCLUDING REMARKS

A harmonised and effective method for the bioinformatic analyses of Whole Genome Sequencing data obtained from both bacteria of importance to human health and SARS-CoV-2 has been developed, and tested by partners in both EU and China.

The use of SNP analysis, as part of this pipeline, allowed for a detailed comparison of these methods, by measuring the accuracy obtained by these laboratories at the level of single nucleotide.

A harmonised approach to manage the bioinformatic analysis of WGS data obtained from food safety relevant bacterial and viral hazards has been implemented.

Through the use of these laboratory methods (described in ***D4.8***) and adoption of a unified bioinformatic pipeline, sequencing applications can now be implemented between all parties on a routine basis.

Furthermore these harmonised methods can also be used to support on-going surveillance across the food chain along with the potential in signalling any changes in epidemiology that may emerge. In this way these data can be shared, if appropriate, to support the risk assessment and risk management of a food safety issue or an outbreak, should this arise and threaten public health and food security.

8. REFERENCES

Bogaerts, B., Winand, R., Fu, Q., Van Braekel, J., Ceyssens, P.-J., Mattheus, W., . . . Vanneste, K. (2019). Validation of a Bioinformatics Workflow for Routine Analysis of Whole-Genome Sequencing Data and Related Challenges for Pathogen Typing in a European National Reference Center: Neisseria meningitidis as a Proof-of-Concept [Original Research]. *Frontiers in Microbiology*, *10*(362). https://doi.org/10.3389/fmicb.2019.00362

Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018 Sep 1;34(17):i884-90.

Croucher NJ, Page AJ, Connor TR, Delaney AJ, Keane JA, Bentley SD, Parkhill J, Harris SR. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. Nucleic acids research. 2015 Feb 18;43(3):e15-.

Land, M., Hauser, L., Jun, S., Nookaew, I., Leuze, M. R., Ahn, T., .Ussery, D. W. (2015). Insights from 20 years of bacterial genome sequencing. *15*(2), 141-161.

Macori, G., Nguyen, S.V., Naithani, A., Hurley, D., Bai, L., El Garch, F., Woehrlé, F., Miossec, C., Roques, B., O'Gaora, P. and Bono, J.L., 2021. Characterisation of Early Positive mcr-1 Resistance Gene and Plasmidome in Escherichia coli Pathogenic Strains Associated with Variable Phylogroups under Colistin Selection. Antibiotics, 10(9), p.1041.

Mitchell, M., Marshall, H., Nguyen, S., Macori, G., & Fanning, S. (2021). The Genomics Revolution: Agri-Food Research in the 21st Century. In *Comprehensive Foodomics* (pp. 2-18): Elsevier.

Quail, M. A., Smith, M., Coupland, P., Otto, T. D., Harris, S. R., Connor, T. R., Gu, Y. (2012). A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *13*(1), 341.

Seemann T. Snippy: fast bacterial variant calling from NGS read. [Internet] https://github.com/tseemann/snippy Available from. Cited 30 December 2018.

Wyres, K. L., Conway, T. C., Garg, S., Queiroz, C., Reumann, M., Holt, K., & Rusu, L. I. (2014). WGS Analysis and Interpretation in Clinical and Public Health Microbiology Laboratories: What Are the Requirements and How Do Existing Tools Compare? *Pathogens*, *3*(2).