EURL-*CAMPYLOBACTER*

REPORT

PROFICIENCY TEST NUMBER 28

Whole genome sequencing of *Campylobacter*

Contents

Abbreviations

AMR	Antimicrobial resistance
C.	*Campylobacter*
EURL	European Union reference laboratory
ISO	International Organization for Standardization
MLST	multi locus sequence typing
cgMLST	core genome MLST
MADe	scaled median absolute deviation
MS	Member State
NGS	next generation sequencing
NRL	national reference laboratory
PT	proficiency test
ST	sequence type
WGS	whole genome sequencing

Introduction

Proficiency test (PT) number 28 on sequencing of *Campylobacter* was sent out by the EU reference laboratory (EURL) for *Campylobacter* in March 2020. The objective of PT 28 was to quantify differences between whole genome sequence (WGS) data from *Campylobacter*, produced at different laboratories. Participation in PT-28 was optional for all national reference laboratories (NRLs).

The results from all participants were compared in terms of different QC parameters for raw sequence data and assembly metrics. To complement this report, individual reports were also prepared for all participants with QC metrics and overall comments on performance.

Outline of the test

The PT contained four samples, including two lyophilised strains of *Campylobacter* and two aliquots of gDNA of the same strains (Table 1).

Participation in PT-28 was optional and 19 NRLs from 14 EU Member States (some Member States have more than one NRL) and Iceland, Norway and Switzerland received the test. Each NRL was given a unique LabID number that was used as an identifier for reporting and uploading of sequence data. LabID will from hereon be L# in this report.

Strain information

The strains used in PT-28, one *Campylobacter jejuni* strain and one *Campylobacter coli* strain, were isolated from broiler chicken samples collected in the *Campylobacter* surveillance program in Sweden. The *C. jejuni* strain was a sequence type (ST)-464 and was isolated in 2016. It had a tetracycline resistance gene (tetO), an arsenite efflux gene (acr3), an organoarsenical efflux gene (arsP), a point mutation (gyrA p.T86I) leading to quinolone resistance, and one (50S_L22_A103V) possibly connected to Macrolide resistance. The *C. coli* strain was a ST-4709 and was isolated in 2017. It had a beta-lactam resistance gene (blaOXA-193 (OXA-61 family class)) (Table 1). The AMR analyses of the reference strains were done by analysing the reference genomes in AMRFinderPlus v.3.2.3 [1], database v. 2019-10-30.1.

Production of reference genomes

The genomes of the two strains were whole genome sequenced using both Illumina MiSeq and Oxford Nanopore Flongle technologies and processed using both Unicycler v0.4.8 [2] and Trycycler v0.3.3 [3]. Hybrid assemblies were generated using both the short-read data from the MiSeq and the long-read data from the Flongle. Complete (gap-free) genomes were obtained for both the *C. jejuni* and the *C. coli* strain with 1,805,160 bp and 1,811,988 bp assembly sizes, respectively. The *C. coli* genome included one plasmid of 104,639 bp.

Table 1. Overview of the different samples used in PT-28.

Sample no.	Sample type	Strain	Isolation year	ST-type	AMR genes/point mutations
PT28-1	gDNA	*Campylobacter jejuni*	2016	464	tetO, acr3, arsP / gyrA p.T86I, 50S_L22_A103V
PT28-3	Lyophilised bacteria				
PT28-2	gDNA	*Campylobacter coli*	2017	4709	blaOXA-193
PT28-4	Lyophilised bacteria				

Sample preparations

Participants received a gDNA sample and a lyophilised bacteria sample from each of the strains, in total four samples (Table 1).

The gDNA samples were prepared from the strains cultivated on horse blood agar. DNA was extracted using a Qiagen EZ1 robot and a Qiagen EZ1 DNA tissue kit (Qiagen, Hilden, Germany). The DNA concentration was measured using a Qubit 2.0 with a DNA dsDNA HS kit (Thermo Scientific, Waltham, MA, USA) and the DNA was stabilised using Biomatrica DNAstable plus solution (Biomatrica, San Diego, CA, USA), which ensured that the DNA was stable in room temperature for the duration of the PT. Each participant received 30 µl of gDNA with a concentration of 13,7 ng/µl from *C. jejuni* and 17 ng/µl from *C. coli*.

The lyophilised cultures used in the PT were produced and tested for stability by the EURL. The vials had a mean bacterial concentration of log 3,54 for the *C. jejuni* strain and log 5,33 for the *C. coli* strain.

Distribution of the proficiency test

The samples for PT-28 were distributed from the EURL on the 9th of March 2020 together with PT-26 and PT-27. The samples were placed in styrofoam boxes along with freezing blocks. The foam boxes were packed in cardboard boxes for transportation and were sent from the EURL using a courier service.

A Micro-T-Log was included in each shipment to record the temperature every second hour during transport.

Procedure

An instruction for the PT was included in the packages and were also sent out by e-mail a few days before the PT distribution. The instruction provided information about storage of the samples and the procedures of the test. The participants were instructed to refrigerate (+2-8°C) the DNA samples and to freeze the vials with lyophilised bacterial cultures (–70°C) on arrival until the DNA sequencing was to be performed.

Participants were instructed to cultivate the strains (samples PT28-3 and PT28-4) and perform DNA extraction according to the procedure normally used at the laboratory. Preparation of sequencing libraries and the sequencing should be performed simultaneously for all 4 DNA samples, using the procedure normally used in the laboratory. Participants were asked to optionally identify the Multi Locus Sequence Type (MLST) as well as the antimicrobial resistance (AMR) genes and/or point mutations causing AMR in the samples. Participants were instructed to upload raw sequence files (FASTQ) without performing any trimming or other modifications prior to submission.

The information about procedures and results had to be reported by each participant in the Questback Essentials system. The link to the survey was sent to all participants on the same day tests were distributed.

Reporting and data submission

The deadline was originally set to the 1st of June 2020 but due to the Covid-19 pandemic the deadline was postponed to the 1st of August 2020. Participants responded to the Questback questionnaire with details about their procedure (the cultivation and DNA extraction procedure, the quality assurance parameters applied, details related to the sequencing and analysis of the obtained sequencing data) and results. Raw sequence files (i.e., FASTQ files) were uploaded by each participant to a personal workspace on the cloud service Onehub (http://www.onehub.com).

Participation

Of the 19 NRLs that registered for participation in PT-28, 11 (58%) reported results through the questionnaire and uploaded sequence data to Onehub before the deadline. Three (16%) of the registered NRLs reported results but failed to upload data before the deadline, whereas 2 (11%) uploaded data but failed to report results. Three (16%) of the registered NRLs did not perform the PT (Figure 1). Only results from the 11 participants that met the deadline for both reporting the results and uploading the data are addressed in this report. One of the 11 participants, L61, did not manage to cultivate the lyophilized bacterial samples and therefore only reported results and uploaded data for the two DNA samples (PT28-1 and PT28-2). The 11 participants represented 10 different EU MS and one third country.

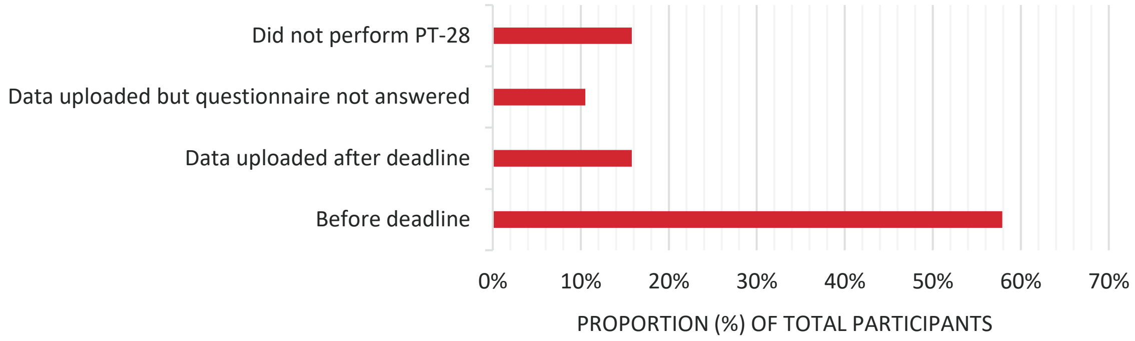

Figure 1. Proportion of participants meeting the deadline for the different tasks of the PT. 19 NRLs signed up for PT-28 participation. 58% met the deadline for results reporting and data uploading, 16% reported results but did not upload data in time, 11% uploaded data but failed to report results and 16% did not perform the PT.

All participants reported that the package had been received without remarks within 3 days of the distribution.

Methods and results reported by participants

Preparation of DNA

Several different techniques were used to extract DNA from the PT28-3 and PT28-4 samples. Seven NRLs used a manual extraction kit method, with the Qiagen QIAamp DNA mini kit being used by 4 NRLs. One NRL used Qiagen MagAttract HMW kit, one used PureLink Genomic DNA minikit and one used A&A Biotechnology Genomic mini kit (Fig. 2). The 4 NRLs that used an automatic procedure each used a different technology; Qiagen QIAcube, Roche MagNA Pure 96, Promega Maxwell 16 and Qiagen EZ1 (Figure 2). All NRLs used a Qubit to measure the DNA concentration except one, which used FLUOstar Omega reader. For quality measurements of the extracted DNA, 7 NRLs used a Nanodrop, one used an Agilent TapeStation and one a DeNovix spectrophotometer. Two NRLs did not assess the quality of the DNA. The quality measured on the Nanodrop (A260/A280 values) were not acceptable according to L24 and L49 for some samples, however L24 tested the DNA integrity on an Agilent Fragment Analyzer system where it was considered acceptable for library preparation. L49 continued to library preparations without any additional measurements.

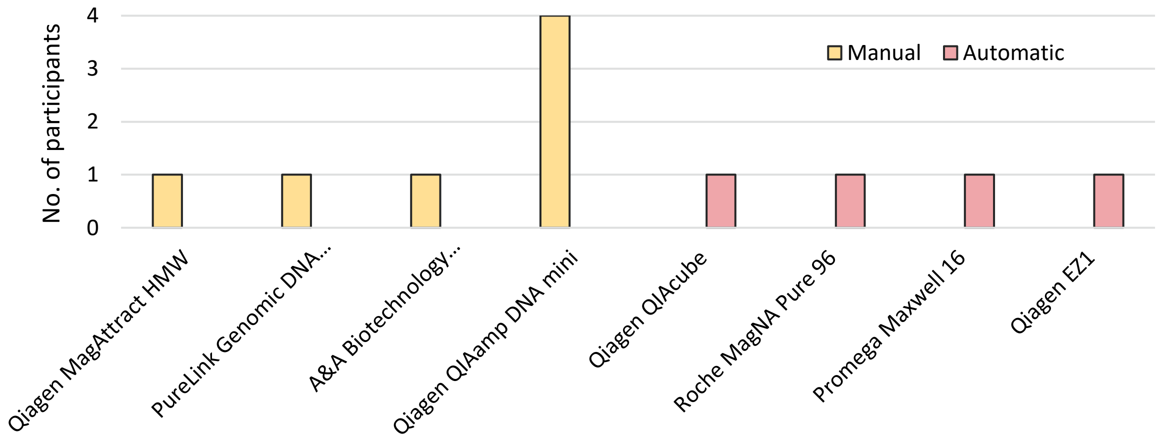

Figure 2. Overview of the different extraction methods used. Four laboratories used the Qiagen QIAamp DNA mini kit to manually extract DNA whereas the remaining 7 laboratories used different methods.

DNA library preparations and sequencing

All the participating NRLs used Illumina technology for their library preparations and sequencing. For library preparations 6 NRLs used the Illumina DNA Prep kit (previously known as Nextera DNA Flex Library Preparation kit) whereas 5 NRLs used the Illumina Nextera XT DNA Library Preparation kit.

For quantification and quality control of the library preparations, 3 NRLs used only Qubit, 2 NRLs used only Agilent Bioanalyzer whereas 2 NRLs used both the Qubit and the Bioanalyzer. One NRL used a Qubit and an Agilent Tapestation and one NRL used only a Tapestation. One NRL then used Fragment Analyser, one used a capillary electrophoresis and one NRL did not perform any quality or quantification control on the prepared library.

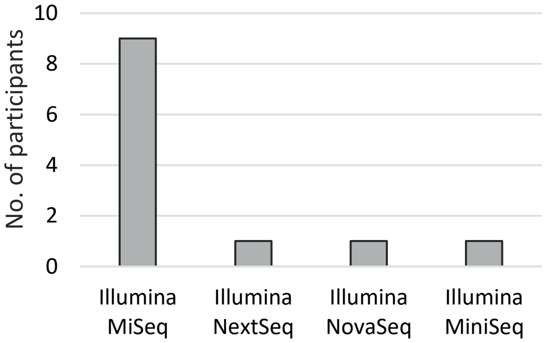

Figure 3. Sequencing instrument used by particpants. Most participating NRLs used an Illumina MiSeq for the sequencing. Illumina NextSeq, Illumina NovaSeq and Illumina MiniSeq were used by one NRL each.

The majority of participating NRLs used Illumina MiSeq for sequencing though L20 used a NovaSeq, L35 used a NextSeq and L58 used a MiniSeq (Figure 3).

MLST analyses

Ten out of the 11 participating NRLs performed the optional MLST analyses. The majority of the NRLs performed the MLST analyses on assemblies but two NRLs did the analyses on raw reads (L23 and 49) (Figure 4).

All the NRLs that performed the MLST analyses could correctly identify the STs for all the samples (*C. jejuni*: ST-464, *C. coli*: ST-4709).

AMR analyses

A total of 9 NRLs performed the optional AMR analyses. Seven of those performed the analyses on assemblies whereas two NRLs did the AMR analyses on raw reads (L23 and L65) (Fig. 4). The participants were instructed to report any genes or point mutations that could possibly lead to AMR.

All the NRLs that performed the AMR analyses were able to identify the tetracycline resistance gene (tetO) for both PT28-1 and PT28-3 (*C. jejuni*) and the quinolone resistance (gyrA. P. T86I) point mutation. Three of the NRLs (L23, L35 and L61) additionally reported a multidrug efflux pump transcription factor gene (cmeR) and L24 reported a point mutation leading to a premature stop codon in the gyrA gene (gyrA p.Q863*). L24 additionally reported a series of point mutations in the cmeR gene. None of the NRLs reported the acr3 and arsP genes and the macrolide resistance point mutation (Table 2 and Table 3).

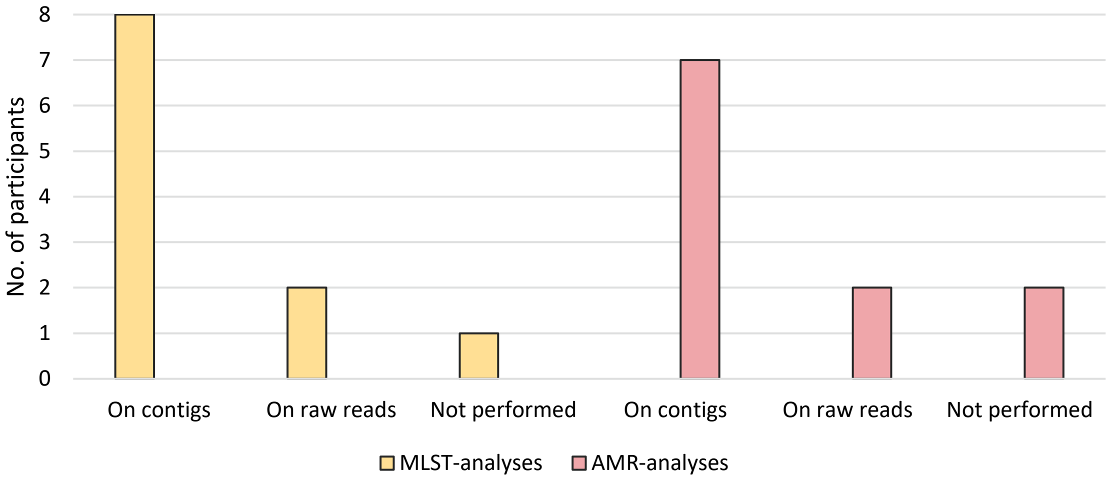

Figure 4. MLST and AMR analysis. Number of participating NRLs that performed MLST and AMR analyses and if the analyses was performed on contigs or raw reads.

For PT28-2 and PT28-4 (*C. coli*) all the laboratories reported the beta-lactam resistance gene (blaOXA) and L35 additionally reported a cmeR gene. L24 reported a low-level streptomycin resistance point mutation (rpsl pA119T) in both the *C. coli* samples. It needs to be noted that one of the NRLs, L61, was not able to cultivate the lyophilised bacteria and did therefore not report results for PT28-3 and PT28-4 (Table 2 and Table 3).

Seven NRLs performed the AMR analyses using ResFinder v3.2. L23 used CARD 3.0.9 in addition to ResFinder. L35 used ABRicate v0.8 and L61 used CARD.

Table 2. The AMR genes detected by the NRLs. All laboratories performing the AMR analyses could identify the tetracycline resistance gene PT28-1 and PT28-3 and the beta-lactam resistance gene in PT28-2 and PT28-4. Additionally, 3 laboratories identified a cmeR multidrug efflux pump transcription factor.

LabID	PT28-1 and PT28-3	PT28-2 and PT28-4
18	tetO	blaOXA-61
19	tetO	blaOXA-61 family gene
23	tetO and cmeR	blaOXA-193 or blaOXA-61 like
24	tetO	blaOXA-like
35	tetO and cmeR	blaOXA-61 and cmeR
49	tetO	blaOXA
58	tetO	blaOXA
61	tetO and cmeR (only PT28-1)	blaOXA-61 (only PT28-3)
65	tetO	blaOXA-193
Ref. strains	tetO – tetracycline resistance acr3 – asenite efflux (stress) arsP – organoarsenical efflux (stress)	blaOXA-193 (OXA-61 family class) – Beta-lactam resistance

Table 3. Point mutations detected by the NRLs that could possibly lead to AMR. All laboratories performing the AMR analyses could identify the point mutation inferring quinolone resistance in PT28-1 and PT28-3. One NRL additionally identified a premature stop codon in the gyrA gene (gyrA p.Q863*). One NRL identified a low-level streptomycin resistance point mutation (rpsL pA119T) in PT28-2 and PT28-4.

LabID	PT28-1 and PT28-3	PT28-2 and PT28-4
18	gyrA p.T86I	None
19	gyrA p.T86I	None
23	gyrA p.T86I ACA>ATA	None
24	gyrA p.T86I – gyrA p.Q863* - cmeR p.T6I – cmeR p.G144D – cmeR p.P183R – cmeR p.S207G	rpsL pA119T GCT>ACT
35	gyrA p.T86I ACA>ATA	None
49	gyrA p.T86I	None
58	gyrA p.T86I	None
61	gyrA p.T86I (only PT28-1)	None (only PT28-3)
65	gyrA p.T86I ACA>ATA	None
Ref. strains	gyrA p.T86I – Quinolone resistance	None

Methods and results from processing of sequence data

Coverage

Most NRLs achieved at least 100X theoretical coverage, except L18, L19, L58 and L61, which had <100X coverage. L18 had 45X coverage for PT28-2, 51X for PT28-3, 63X for PT28-1 and 72X for PT28-4. L19 had 55X coverage for PT28-3 whereas other samples had 80X-85X coverage. The coverage for L58 was 66-76X for all the samples and L61 had only 53X for PT28-1 and 45X for PT28-2 (Table 5 and Figure 11).

Quantification of high-quality bases

The number of bases with quality values above Q30 in each FASTQ-file was quantified by an in-house made Perl script. The percentage of bases with a quality score of at least 30 is presented in Figure 5. The median value was 90% and the range was from 75%-97%. Many laboratories had a lower quality of bases in the R2 reads.

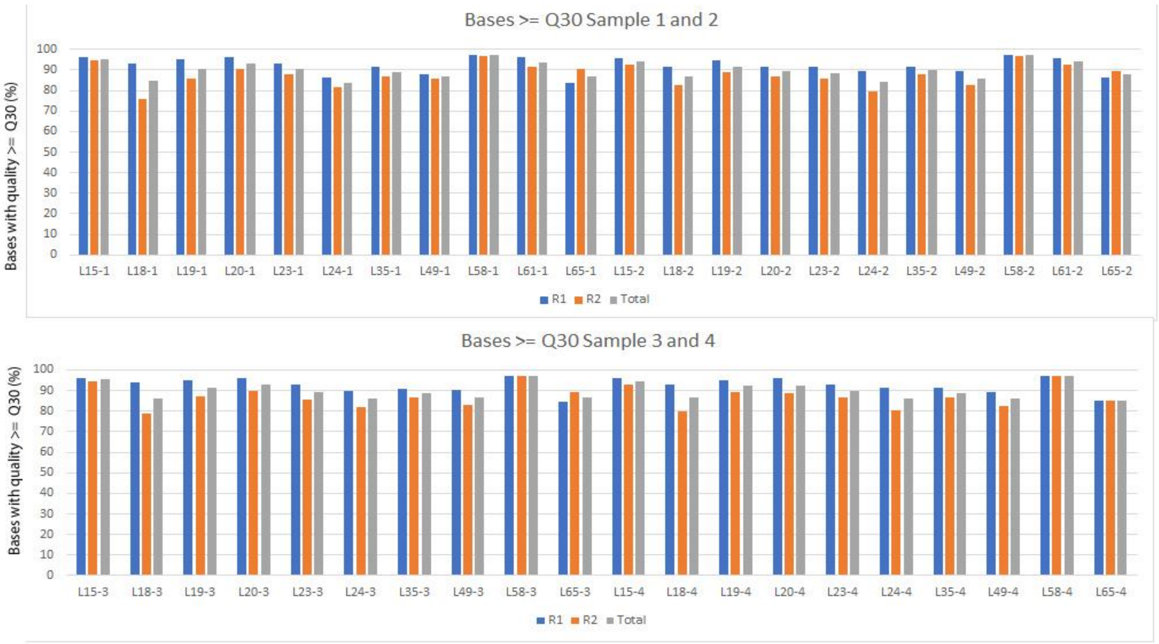

Figure 5. Percentage Q30 bases. Shows the percentage of bases with quality values of >= Q30.

Trimming analysis

Reads were trimmed with Trimmomatic 0.39 [5] using the same parameters as in the assembly pipeline, but in two steps. First adapters were removed, then low quality bases were trimmed (LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36). The number of reads and bases were quantified before and after each trimming step using an in-house made perl script. Three laboratories (L20, L24 and L49) had between 8 and 17% bases related to adapter sequences. Actually, laboratories were instructed not to perform any trimming of the data, which can have been interpreted by some participants to not let the sequencer perform the adapter-removal. The amount of low-quality bases trimmed ranged from 1 to 16%. L18, L19 and L65 had over 10% quality trimmed bases in at least one of the samples. See Table 4 for the trimming results.

Sequence contamination

The sequence data (FASTQ-files) was processed with the Kraken2 [4] software to obtain metagenomic information about the sequencing datasets. Kraken2 classifies reads as belonging to different phylogenetic taxa and this indicates if the correct species was sequenced and if the samples contained contaminating reads from a different organism. The most important finding was that sample PT28-3 from L20 was a *C. coli* when classifying the reads, which indicates a mix-up of the samples. The *C. coli* genome present in those read-files was analysed and it was the same strain as used in this PT (represented by samples PT28-2 and PT28-4). This sample was excluded for further analysis.

Table 4. Results of the trimming analyses. The table shows a summary of the FASTQ-files and results from the trimming analyses.

Sample	Lab	Readpairs (millions)	Readlength	Bases (millions)	Reported Trimmed by submitter	Bases lost due to adapter trimming (%)	Bases lost due to quality trimming (%)	Readpairs surviving (millions)	Bases surviving (millions)	Bases surviving (percent)	Theoretical coverage (X)
Sample1	L15	0,56	250	267,0		0,1	1,5	0,56	263	98,4	148
	L18	0,27	300	135,7		0,6	16,5	0,25	113	83,0	63
	L19	0,54	150	160,4	yes	0,1	11,2	0,49	142	88,7	80
	L20	8,14	150	2457,2		15,8	2,6	7,98	2015	82,0	1134
	L23	1,10	300	586,9		0,3	5,9	1,07	551	93,8	310
	L24	0,96	300	577,5	yes	12,5	5,9	0,95	476	82,4	268
	L35	2,34	150	695,9	yes	0,8	5,1	2,27	655	94,1	369
	L49	0,50	300	302,0	yes	12,6	4,4	0,49	252	83,6	142
	L58	0,46	150	138,8		0,9	1,5	0,46	135	97,6	76
	L61	0,31	250	95,9		0,3	1,9	0,31	94	97,7	53
	L65	1,10	300	468,8	yes	1,3	8,0	1,08	426	90,9	240
Sample2	L15	0,54	250	257,7		0,1	2,1	0,54	252	97,9	139
	L18	0,23	300	92,4		1,1	10,8	0,22	81	88,2	45
	L19	0,58	150	168,6	yes	0,1	8,3	0,54	155	91,7	85
	L20	4,24	150	1281,3		15,8	3,9	4,13	1037	80,9	572
	L23	0,95	300	515,5		0,3	7,1	0,93	477	92,6	264
	L24	1,31	300	786,0	yes	13,5	6,4	1,28	636	81,0	351
	L35	2,23	150	665,3	yes	0,9	4,6	2,18	629	94,6	347
	L49	0,61	300	369,8	yes	14,2	5,7	0,60	299	80,9	165
	L58	0,45	150	137,3		1,0	1,4	0,45	134	97,6	74
	L61	0,24	250	82,4		0,3	1,8	0,23	81	97,9	45
	L65	0,74	300	300,0	yes	0,9	7,6	0,73	275	91,6	152
Sample3	L15	0,68	250	319,9		0,1	1,5	0,67	315	98,4	177
	L18	0,25	300	106,5		1,1	14,1	0,23	90	84,9	51
	L19	0,37	150	107,3	yes	0,1	9,7	0,34	97	90,2	55
	L20	4,67	150	1410,9		17,5	2,3	4,58	1137	80,6	640
	L23	0,91	300	502,7		0,4	7,3	0,88	464	92,3	261
	L24	1,19	300	717,5	yes	13,9	5,0	1,17	587	81,8	331
	L35	1,12	150	333,4	yes	1,2	5,2	1,09	312	93,6	176
	L49	0,67	300	404,1	yes	12,8	5,6	0,65	333	82,3	187
	L58	0,40	150	121,3		1,6	1,4	0,40	118	97,1	66
	L65	0,94	300	421,1	yes	0,9	8,3	0,93	382	90,8	215
Sample4	L15	0,62	250	293,8		0,1	1,9	0,62	288	98,0	159
	L18	0,34	300	150,2		0,9	12,8	0,32	130	86,4	72
	L19	0,54	150	159,3	yes	0,1	7,4	0,51	147	92,5	81
	L20	4,54	150	1372,0		15,5	2,9	4,44	1125	82,0	621
	L23	1,07	300	593,3		0,3	6,6	1,05	552	93,1	305
	L24	0,70	300	422,9	yes	8,1	8,1	0,68	357	84,5	197
	L35	1,98	150	590,6	yes	0,7	5,2	1,92	556	94,1	307
	L49	0,70	300	421,2	yes	11,6	6,0	0,68	350	83,1	193
	L58	0,47	150	141,1		1,0	1,4	0,46	138	97,6	76
	L65	0,84	300	434,8	yes	0,4	10,2	0,82	389	89,4	215

The most common contaminating organism was *Alteromonas macleodii* and reads from it were found in sequencing data from 5 NRLs in all four samples. Reads from *Listeria* were found in samples from 3 NRLs. L19 had *Vibrio cholerae* in all the samples and L65 had *E. coli* in all the samples. *Salmonella* was found in samples from two NRLs, L23 and L61, and one NRL, L15 had *Klebsiella pneumoniae* in the PT28-3 sample. No participant had all the samples completely free from contamination. The levels of contamination were very low, with the highest being 1,87% representing *A. macleodii* in sample PT28-2 from L24. Such low-level contamination usually does not interfere with the assembly process and downstream analyses but may have an impact on AMR determination and other defined characteristics (Table 5).

Table 5. Contaminations in sequencing data. A list of the most common contaminations found in the sequencing data, processed with Kraken2.

LabID	PT28-1	PT28-2	PT28-3	PT28-4
15	0,10% *E. coli*	0,08% *E. coli*	0,04% *K.pneumoniae*	0,19% *E. coli*
18	0,03% *L. monocytogenes*	0,07% *L. monocytogenes* 0,06% *Neisseria*	0,06% *L. monocytogenes* 0,0% *Neisseria*	0,07% *L. monocytogenes*
19	0,12% *Vibrio cholerae*	0,09% *Vibrio cholerae*	0,11% *Vibrio cholerae*	0,11% *Vibrio cholerae*
20	0,02% *A. macleodii*	0,02% *A.macleodii*	0,02% *A.macleodii*	0,04% *A. macleodii*
23	No contaminaton	No contamination	No contamination	0,01% *S.enterica*
24	*1,07% A. macleodii* 0,09% *L. monocytogenes*	*1,87% A. macleodii* 0,06% *L. monocytogenes*	*1,21% A. macleodii* 0,07% *L. monocytogenes*	*0,88% A. macleodii* 0,05% *L. monocytogenes*
35	0,07% *A. macleodii*	0,06% *A. macleodii*	0,21% *A. macleodii*	0,07% *A. macleodii*
49	0,77% *A. macleodii*	1,65% *A. macleodii*	0,95% *A. macleodii*	0,92% *A. macleodii*
58	0,19% *A. macleodii*	0,21% *A. macleodii*	0,31% *A. macleodii*	0,22% *A. macleodii*
61	0,12% *Listeria* 0,05% *Salmonella*	0,14% *Listeria* 0,06% *Salmonella*	No sample	No sample
65	0,03% *E. coli*	0,05% *E. coli*	0,02% *E. coli*	0,01% *E. coli*

Overall alignment rate to reference genomes

Trimmed reads were aligned to each of the reference genomes using Bowtie2 [6] with default settings. The number of aligned reads was counted. All laboratories yielded a high overall alignment rate, ranging from 98.3 to almost 99.9% (Fig. 6).

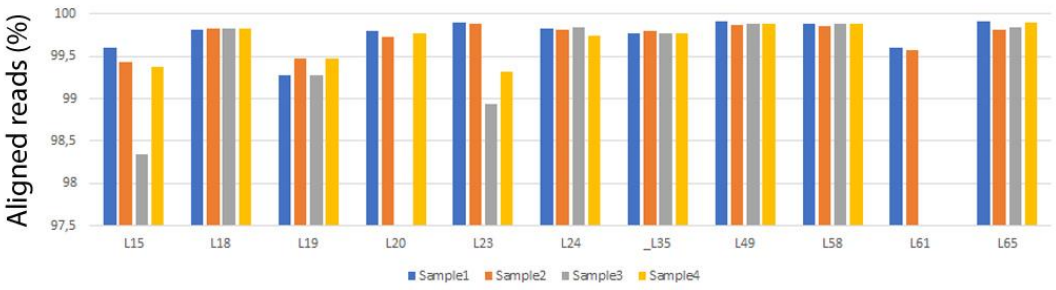

Figure 6. Overall alignment rate. Shows the proportion of reads that could be mapped to the corresponding reference genome.

Insert Fragment size

The insert sizes were defined by comparison of the distances between the forward and reverse read mapping positions in the reference genomes. Most laboratories had fragment sizes around 300 bp (Figure 7). L20, L65, L18 and L61 had fragments shorter than 200 bp in at least one sample, which can lead to loss of sequence efficiency due to sequencing of adapters that have to be trimmed away.

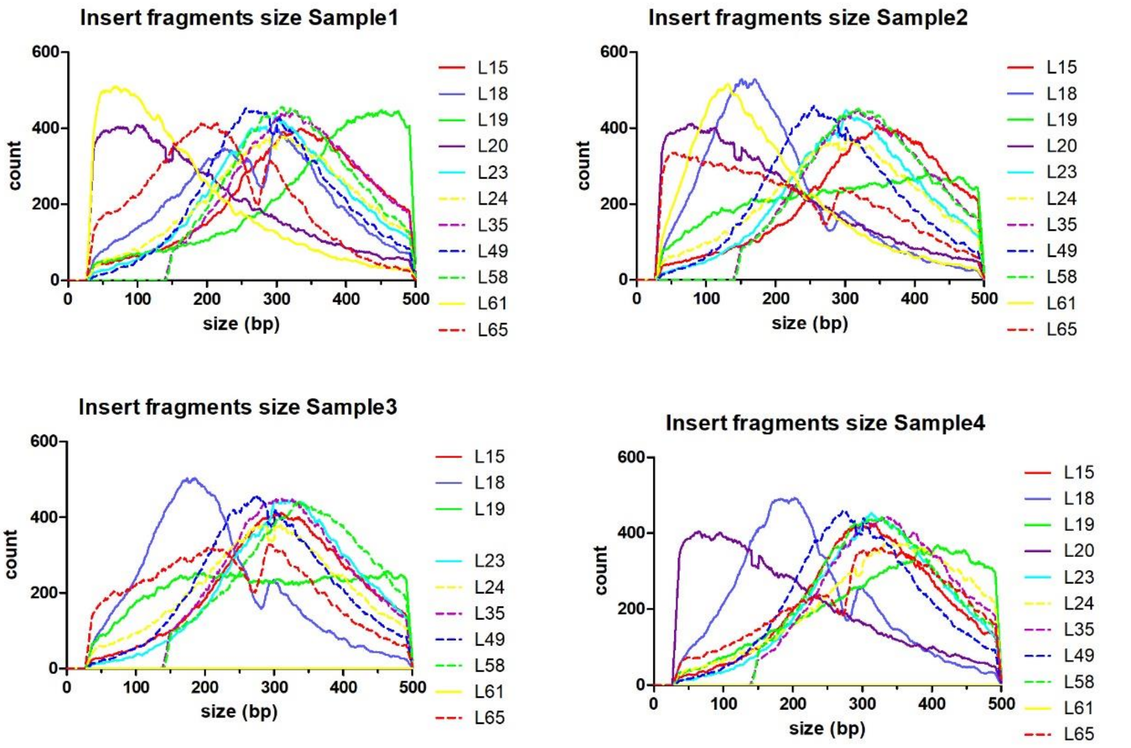

Figure 7. Inserts/fragments size. The insert/fragment sizes that were sequenced were determined by mapping the positions of the forward and reverse reads in the reference genomes. 500 bp is the maximum value possible when mapping with Bowtie2 which explains the drop at 500 bp for all samples.

Distribution of reads over the reference genomes

The percentage of each reference genome that was covered by at least one read using mapping data corresponding to sequence depths between 10X and 100X was calculated and is shown in Figure 8. The alignment file was down-sampled to contain reads corresponding to 10, 20, 30, 40, 50, 60, 70, 80, 90 and 100X sequence depth and the percentage of the reference genome covered by at least one read was quantified using a SAMtools [7] mpileup and in-house made scripts. L65 had very uneven distribution of mapped reads with only 95% of the reference genome covered at 100X sequence depth. L18, L20, L61, L24 had more uneven distribution as compared to L15, L19, L23, L35, L49 and L58. PT28-4 showed less evenly distributed reads for L15 than the other three samples. At 100X sequence depth, all samples were covered by all participants, except L65, at basically all positions in the reference genomes with at least one sequence read.

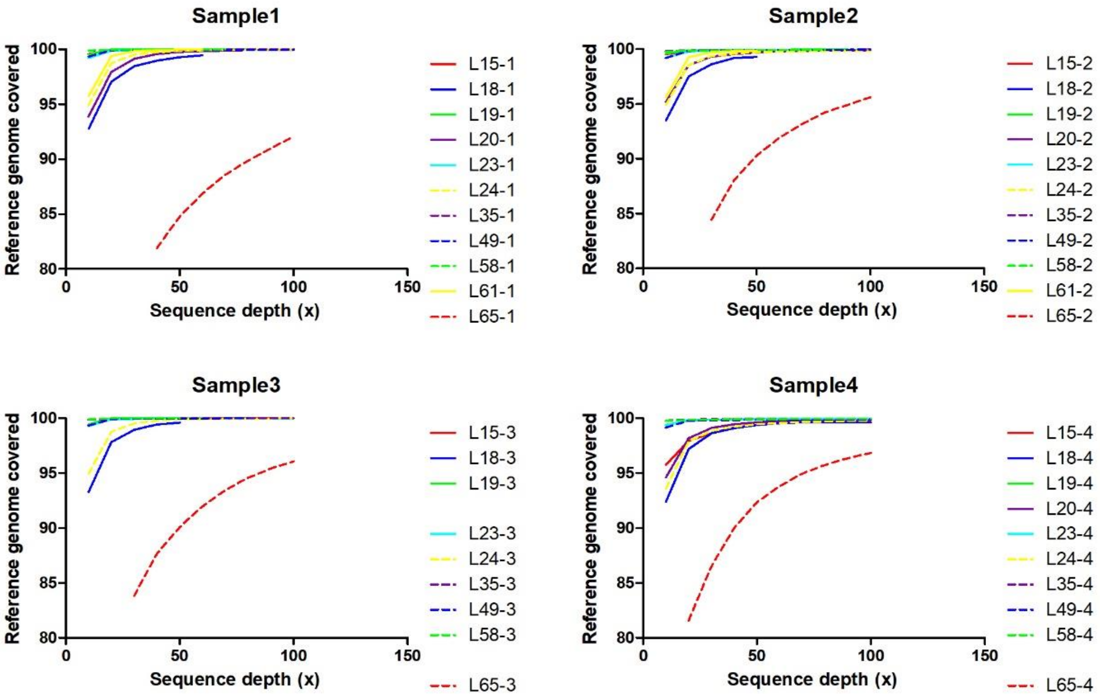

Figure 8. Distribution of mapped reads over the reference genomes. This figure shows the percentages of the reference genomes that were covered by at least one read using mapping data corresponding to sequence depths between 10 and 100X coverage.

Assemblies

Reads were quality trimmed using Trimmomatic 0.39 [5], downsampled to 100x coverage, followed by assembly with the SPAdes software 3.14.1 [8]. The assemblies were error corrected using the Pilon software [9].

The following QC metrics were calculated for each assembly:

- Total size of assembly (bp)
- GC content
- Total number of contigs
- Total number of contigs > 1kb
- Longest contig
- N50 length

All the submitted raw sequencing data could be assembled, except for sample PT28-4 from L15. That data did not assemble due to mismatches in the naming between the R1 and R2 reads in a small fraction of the read pairs. This appeared to be due to a small displacement of the coordinates of the clusters between the R1 and R2 reads and affected less than 1,000 read pairs out of 620,000. Therefore, the error-correcting step included in SPAdes could not be performed. Additionally, L61 did not upload data for PT28-3 and PT28-4 as described previously. The QC metrics for each assembly is summarised in Appendix A.

Total number of contigs

In this analysis, no filtering or cut-off was applied. Most of the data assembled into less than 100 contigs, but some generated more than 100 contigs and data from three participants generated more than 200 contigs (Figure 9). Many of those contigs were from contaminating reads. One *C. jejuni* assembly for L15 had over 250 contigs, all assemblies except PT28-3 for L19 had above 200 contigs, and all the assemblies for L65 were very fragmented, ranging from 302 for PT28-4 to 643 contigs for PT28-1.

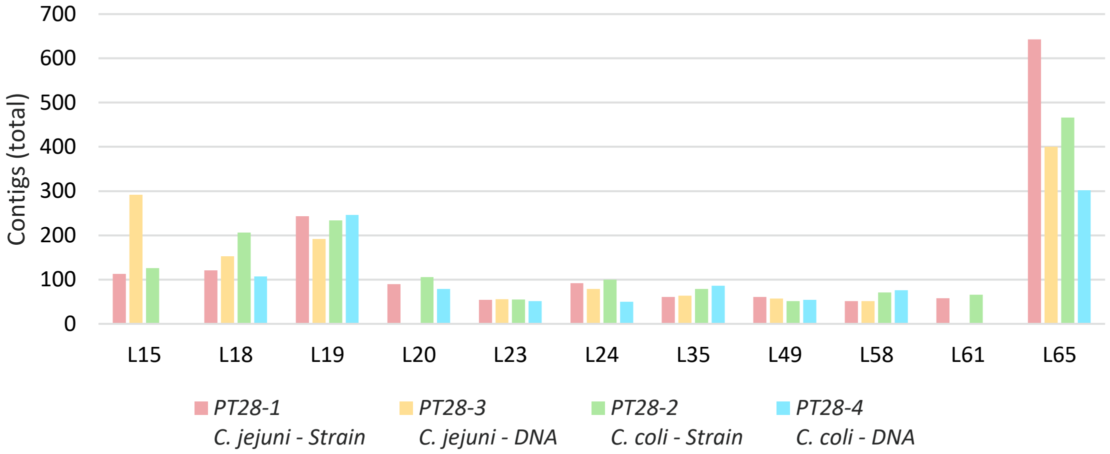

Figure 9. Number of contigs. The total number of contigs for all the assembled sequencing data.

N50 length

The largest N50 lengths were obtained for the *C. coli* samples (PT28-2 and PT28-4). Many of the *C. coli* assemblies had N50 lengths around 200 kb. The lowest N50 lengths were from L18 and L65, where all the assemblies had very low N50 length (<62.5 kb for L18 and <12 kb for L65) (Figure 10).

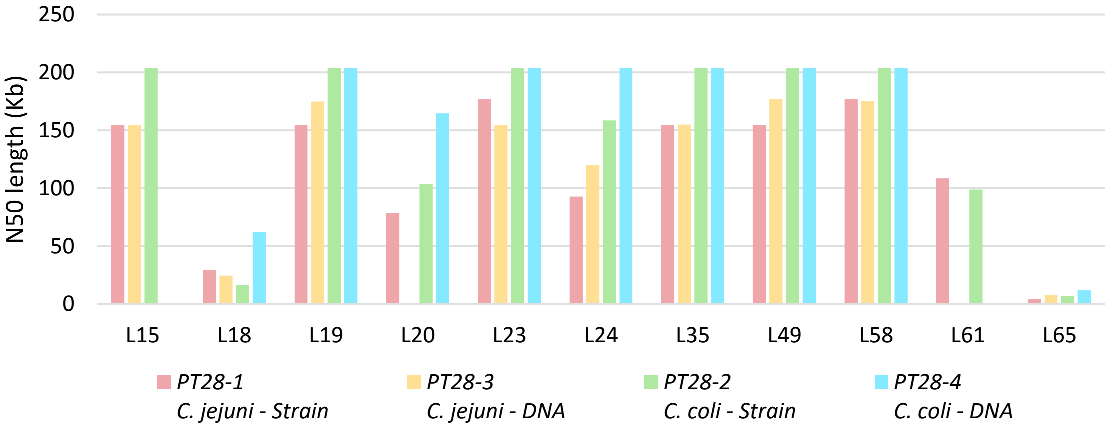

Figure 10. The N50 lengths for all the assembled sequencing data.

Assembly size

The sizes of the PT28-2 and PT28-4 (*C. coli*) assemblies were mostly in an accordance with the size of the reference genome. The *C. jejuni* reference genome had a ~40 kb duplicated region, which was not resolved with short read sequencing. This explains why most PT28-1 and PT28-3 assemblies were smaller in size than the *C. jejuni* reference genome. One outlier was the PT28-3 assembly from data from L15, which was substantially larger than the reference genome. That assembly produced almost 300 contigs and only 36 of them were more than 1 kb in size so the multitude of small contigs is one possible cause to the large assembly size. Many of the short contigs were from *Klebsiella pneumoniae* (0,04 % of the total number of reads were classified as being from *K. pneumoniae*). L18 and L65 had the smallest assembly sizes for both the *C. jejuni* samples, and the *C. coli* samples. L65 assemblies were substantially smaller than the reference genomes generating assembly sizes of 1,502,770 bp and 1,644,818 bp for PT28-1 and PT28-3, respectively and 1,683,691 bp and 1,719,654 bp for PT28-2 and PT28-4, respectively (Figure 11). In addition, L65 had high variations between the corresponding assemblies for both the *C. jejuni* samples (142,048 bp difference) and the *C. coli* samples (35,963 bp difference). Other participants had low variations between the corresponding assemblies. (Figure 11).

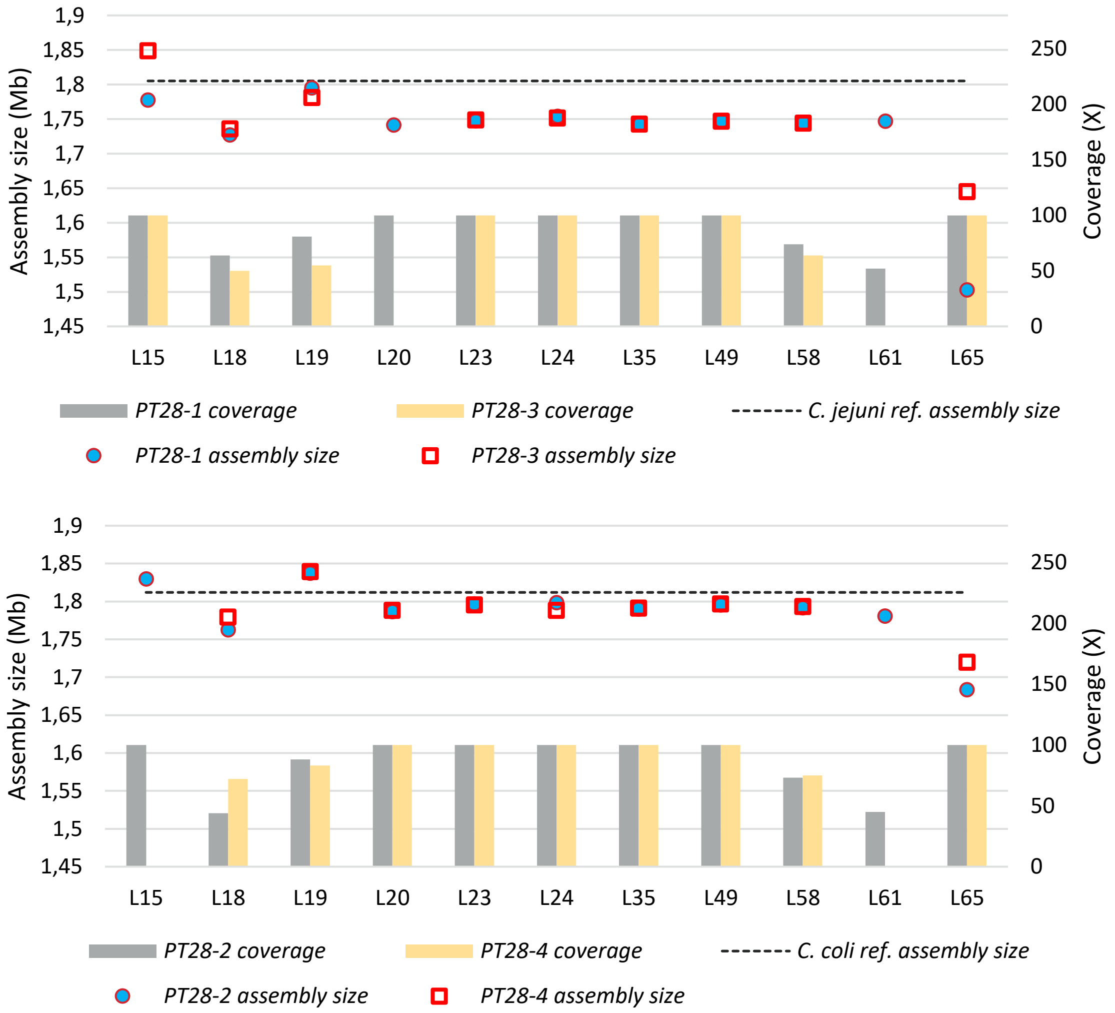

Figure 11. Sequencing coverage and assembly sizes. The primary axis shows the assembly sizes with dotted lines representing the references strains and the secondary axis shows the sequencing coverage. Top: *C. jejuni* samples, PT28-1 and PT28-3, bottom: *C. coli* samples, PT28-2 and PT28-4.

Allele calling and clustering

The assemblies were imported into the SeqSphere+ software (Ridom GmbH) using the core genome scheme Oxford v.1 [10] for *C. jejuni* and an ad hoc core genome MLST for *C. coli* created by the EURL (Seed genome: NZ_CP023545. Query genomes: CP011015, CP007183, CP017873. Default settings, SeqSphere+). The number of recovered alleles were used as an additional QC metric for each assembly.

Only 4 *C. jejuni* assemblies recovered less than 97% of the targets. L18 recovered 94,8% for both PT28-1 and PT28-3, and L65 recovered 79,1% for PT28-1 and 80,8% for PT28-3. For *C. coli*, all assemblies recovered more than 97% of the targets except L18 and L65. L18 assemblies still recovered a high proportion of targets (93,5% for PT28-2 and 97% for PT28-4) whereas L65 recovered substantially fewer targets (79,1% for PT28-3 and 88,9% for PT28-4) (Figure 12).

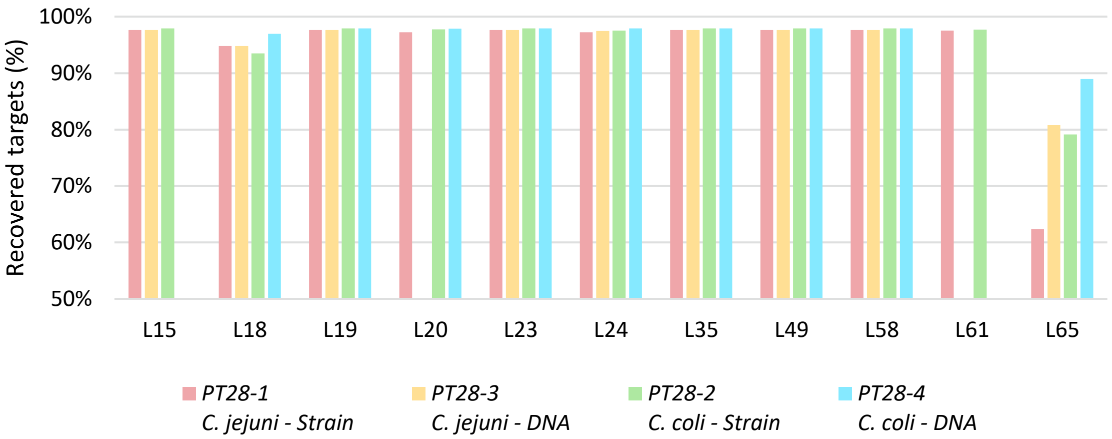

Figure 12. The proportion of recovered targets using species specific cgMLST schemes. The assemblies were imported into SeqSphere+ either with the Oxford v.1 scheme (*C. jejuni*, 1343 targets) or an ad hoc cgMLST scheme (*C. coli*, 1121 targets).

A cluster analyses of the *C. jejuni* samples based on the Oxford v.1 scheme revealed only one allele difference between all samples, except samples from L18 and L65. L18 samples only showed 3-4 allele differences from the main cluster, but the L65 samples had substantially larger differences (Figure 13A). The cluster analyses for the *C. coli* samples based on the ad hoc cgMLST scheme displayed similar results as for the *C. jejuni* samples, except that the L18 PT28-4 sample only showed one allele difference from the main cluster (Figure 13B).

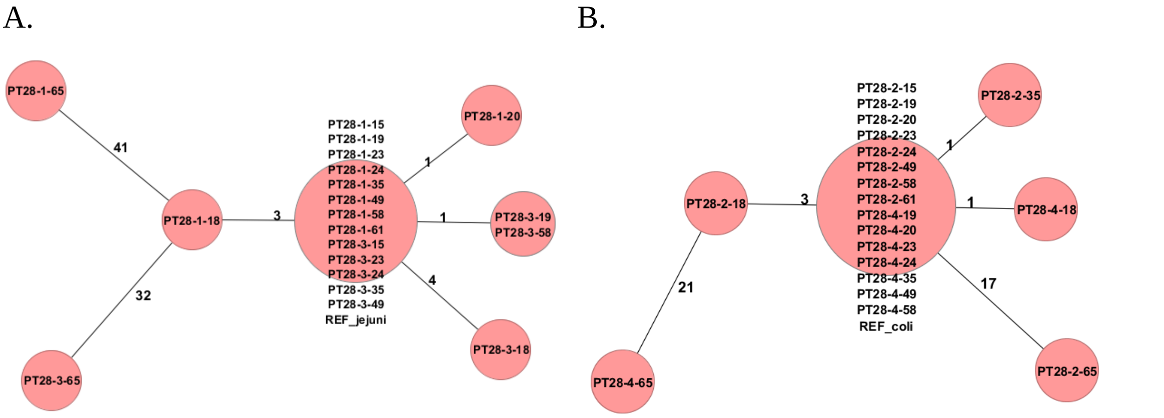

Figure 13. MST figures based on cgMLST cluster analyses. A) *C. jejuni* assemblies analyzed using the Oxford v.1 scheme with 1343 targets. The reference genome is included as REF_jejuni. **B)** *C. coli* samples analyzed using an ad hoc cgMLST scheme with 1121 targets. The reference genome is included as REF_coli.

Performance thresholds

Performance thresholds were defined for assembly size (lower and upper value), N50 length and total number of contigs by calculating the median value ±3σMADe for each sample according to ISO 22117:2019 (Table 6) [11]. A scoring system was not defined for PT-28. However, values outside median values ±3σMADe were defined as outliers and should be considered as indications that performance can be improved. However, values outside the thresholds do not necessarily implicate a poor performance since data analysis may be optimised in each laboratory and include e.g. filtering or cut-off values in assemblies. Overall comments on the data and possible focus areas for improving performance were commented further in each laboratory's individual report.

Table 6. PT-28 performance thresholds. Performance thresholds were defined for three quality parameters for each sample (assembly size, number of contigs and N50 length.

Sample	Quality parameter	Median value	Median+3σMADe	Median-3σMADe
PT28-1	Assembly size (bp)	1746750	1771862	1721638
	Number of contigs	90	228	
	N50 length (bp)	154573		55187
PT28-2	Assembly size (bp)	1792001	1821864	1762138
	Number of contigs	100	251	
	N50 length (bp)	203647		202762
PT28-3	Assembly size (bp)	1746722	1781877	1711567
	Number of contigs	90	192	
	N50 length (bp)	154717		59554
PT28-4	Assembly size (bp)	1791117	1810896	1771338
	Number of contigs	90	204	
	N50 length (bp)	203691		203002

Summary of proficiency test number 28, 2020

The objective of PT-28 was to quantify differences between whole genome sequence (WGS) data from *Campylobacter*, produced at different laboratories, with the purpose to harmonise the production of reliable laboratory results. Due to the Covid-19 pandemic, the participants were fewer than expected, with 11 out of 19 participants fulfilling the requirements of submitting both the survey and data before the extended deadline.

Overall, the test showed that most laboratories performed very well. Six laboratories: L15, L18, L19, L20, L61 and L65 were defined as outliers for more than one quality parameters or samples, or as a result of missing data.

Outlying results does not necessarily implicate poor performance in downstream analysis such as AMR, ST definition or allele calling. Only data from L18 and L65 deviated with more than one allele in the cgMLST analysis. L18 could probably have improved the performance with a higher sequencing depth, whereas L65 needs to address technical issues resulting in a very uneven distribution of reads along the reference genomes.

References

[1] M. Feldgarden *et al.*, “Validating the AMRFinder Tool and Resistance Gene Database by Using Antimicrobial Resistance Genotype-Phenotype Correlations in a Collection of Isolates,” *Antimicrob Agents Chemother*, vol. 63, no. 11, pp. e00483-19, /aac/63/11/AAC.00483-19.atom, Aug. 2019, doi: 10.1128/AAC.00483-19.

[2] R. R. Wick, L. M. Judd, C. L. Gorrie, and K. E. Holt, “Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads,” *PLOS Computational Biology*, vol. 13, no. 6, p. e1005595, Jun. 2017, doi: 10.1371/journal.pcbi.1005595.

[3] R. Wick, *rrwick/Trycycler: Trycycler v0.4.1*. Zenodo, 2020.

[4] D. E. Wood, J. Lu, and B. Langmead, “Improved metagenomic analysis with Kraken 2,” *Genome Biology*, vol. 20, no. 1, p. 257, Nov. 2019, doi: 10.1186/s13059-019-1891-0.

[5] A. M. Bolger, M. Lohse, and B. Usadel, “Trimmomatic: a flexible trimmer for Illumina sequence data,” *Bioinformatics*, vol. 30, no. 15, pp. 2114–2120, Aug. 2014, doi: 10.1093/bioinformatics/btu170.

[6] B. Langmead and S. L. Salzberg, “Fast gapped-read alignment with Bowtie 2,” *Nature Methods*, vol. 9, no. 4, Art. no. 4, Apr. 2012, doi: 10.1038/nmeth.1923.

[7] H. Li *et al.*, “The Sequence Alignment/Map format and SAMtools,” *Bioinformatics*, vol. 25, no. 16, pp. 2078–2079, Aug. 2009, doi: 10.1093/bioinformatics/btp352.

[8] A. Bankevich *et al.*, “SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing,” *J Comput Biol*, vol. 19, no. 5, pp. 455–477, May 2012, doi: 10.1089/cmb.2012.0021.

[9] B. J. Walker *et al.*, “Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement,” *PLOS ONE*, vol. 9, no. 11, p. e112963, Nov. 2014, doi: 10.1371/journal.pone.0112963.

[10] A. J. Cody, J. E. Bray, K. A. Jolley, N. D. McCarthy, and M. C. J. Maiden, “Core Genome Multilocus Sequence Typing Scheme for Stable, Comparative Analyses of *Campylobacter jejuni* and *C. coli* Human Disease Isolates,” *Journal of Clinical Microbiology*, vol. 55, no. 7, pp. 2086–2097, Jul. 2017, doi: 10.1128/JCM.00080-17.

[11] ISO/TS 22117:2019: Microbiology of food and animal feeding stuffs – Specific requirements and guidance for proficiency testing by interlaboratory comparison. International Organization for Standardization.

Appendix A

Summary of all QC metrics for the assembled NRL data.

LabID	Sample	Assembly size (kb)	GC%	Contigs total	Contigs >1kb	Longest contig (kb)	N50 (kb)
15	**PT28-1**	1777139	30,5	113	30	287285	154717
	PT28-2	1829653	31,45	126	27	352829	203846
	PT28-3	1848302	31,76	292	36	287285	154717
18	**PT28-1**	1726994	29,98	121	101	143833	29113
	PT28-2	1762737	31,11	206	167	67162	16506
	PT28-3	1735322	30,4	153	119	79782	24460
	PT28-4	1779227	31,39	107	78	132512	62457
19	**PT28-1**	1794734	30,99	243	28	287185	154617
	PT28-2	1837565	31,45	234	32	352379	203691
	PT28-3	1780636	31,19	192	30	287184	174979
	PT28-4	1839167	31,47	246	31	352379	203691
20	**PT28-1**	1741104	29,93	90	54	144747	78524
	PT28-2	1786885	31,17	106	50	255942	103752
	PT28-4	1788325	31,2	79	41	278661	164732
23	**PT28-1**	1747144	30,28	54	32	287285	176918
	PT28-2	1797024	30,74	55	27	391185	203846
	PT28-3	1748412	29,98	56	30	287285	154717
	PT28-4	1795564	31,04	51	29	352829	203846
24	**PT28-1**	1753675	30,6	92	40	287241	92685
	PT28-2	1798715	31,27	100	41	352829	158662
	PT28-3	1751130	29,85	79	35	287285	119798
	PT28-4	1788238	31,62	50	32	278661	203846
35	**PT28-1**	1742313	30,47	61	29	287141	154573
	PT28-2	1790040	30,16	79	31	391041	203647
	PT28-3	1742550	30,55	64	29	287141	154893
	PT28-4	1791117	30,45	86	31	352335	203647
49	**PT28-1**	1747012	30,27	61	35	287285	154717
	PT28-2	1795830	30,67	51	27	391185	203846
	PT28-3	1746722	30,06	57	33	287254	176949
	PT28-4	1796564	30,97	54	26	352829	203846
58	**PT28-1**	1744082	28,89	51	30	287185	176818
	PT28-2	1792001	30,53	71	30	352379	203746
	PT28-3	1743769	31,03	51	28	325425	175276
	PT28-4	1793270	30,91	76	31	352652	203746
61	**PT28-1**	1746750	30,29	58	35	287182	108454
	PT28-2	1780717	31,4	66	44	210794	98801
65	**PT28-1**	1502770	32,36	643	413	20794	3974
	PT28-2	1683691	32,32	466	308	25141	7107
	PT28-3	1644818	31,1	400	282	23903	7906
	PT28-4	1719654	32,15	302	215	57873	11904