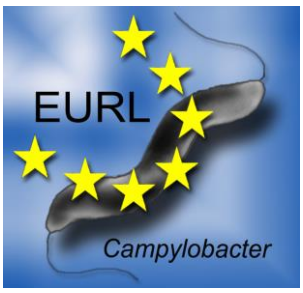


EURL-*CAMPYLOBACTER*

REPORT

PROFICIENCY TEST NUMBER 25

Subtyping of *Campylobacter jejuni* using MLST or WGS

Contents

Abbreviations

C.	*Campylobacter*
EURL	European Union reference laboratory
MLST	multi locus sequence typing
cgMLST	core genome MLST
wgMLST	whole genome MLST
MS	member state
NGS	next generation sequencing
NRL	national reference laboratory
PT	proficiency test
SNP	single nucleotide polymorphism
ST	sequence type
WGS	whole genome sequencing

Introduction

Proficiency test (PT) number 25 on subtyping of *Campylobacter jejuni* using MLST or WGS was organised by the EU reference laboratory (EURL) for *Campylobacter* in March 2019. The objective of the PT was to assess the performance of subtyping eight strains of *C. jejuni* that were delivered as extracted DNA to the participants. The multi locus sequence type (ST) was to be determined for each strain by the participating laboratories and the maximum score was therefore eight. No performance thresholds were set for this PT.

To acquire the ST of a bacterial isolate, parts of the DNA-sequences of seven house-keeping genes need to be determined and compared to other previously sequenced genes. The instances of each gene fragment are called alleles and each allele is given a unique allele identifier when queried to an allele identifier database. For *Campylobacter* spp. and many other pathogenic bacteria, the PubMLST database (pubmlst.org) is the authority and sequenced alleles needs to be queried to that database or an instance of it via different software. The combination of the seven alleles yields a sequence type (ST) determined by the database, and these were to be reported to the EURL along with answers to questions regarding the methods used.

To acquire the DNA-sequences, the participants could use either Sanger-based sequencing (in which only the seven gene fragments are sequenced) or chose to sequence the whole genome of the sample strains. Participants opting for the whole genome sequencing (WGS) approach were also asked to perform a relatedness analysis and to identify clusters of closely related strains. This will be referred to as a cluster analysis. The creation of a fictional scenario accompanied the PT in order to mimic an actual outbreak investigation. The cluster analysis was optional and was not scored. However, the participants' interpretations of the clusters along with analysis performed by the EURL on the submitted data is discussed in this report.

Outline of the proficiency test

Selection of strains

The eight strains had been isolated from chicken samples collected within the surveillance program in Sweden during 2014-2015 or from samples from wild birds collected during an investigation in the Antarctic 2012. The eight strains represented five different STs (Table 1).

Table 1. The eight strains used in PT 25 and their sequence types (ST).

Sample ID	Sequence Type (ST)
PT25-1	257
PT25-2	21
PT25-3	21
PT25-4	257
PT25-5	883
PT25-6	257
PT25-7	1326
PT25-8	45

To enable a cluster analysis, some strains had to be relatively closely related to each other, thus belonging to the same ST. The aim was that the strains would produce two clusters in cluster analysis – one with an easy interpretation and one that posed a bigger challenge in interpretation. The samples that would produce the "easy cluster" were the same DNA put in two different tubes and marked with different sample numbers. The more difficult cluster contained three strains, in which two were almost inseparable (but originated from different birds at different locations) and a third strain that was a little bit more distantly related. The strains were sequenced by the EURL using both Sanger-based sequencing and WGS and the correct STs could be determined for each strain with both techniques. For WGS, the sequence reads were quality trimmed using Trimmomatic 0.36 [1] followed by an assembly with the SPAdes software 3.13.0 [2]. The assemblies were error corrected using the Pilon software [3] and the resulting assemblies were analysed in SeqSphere+ software (Ridom GmbH) using the core genome scheme Oxford v.1 [4]. The core genome MLST (cgMLST) analysis is visualised in Figure 1.

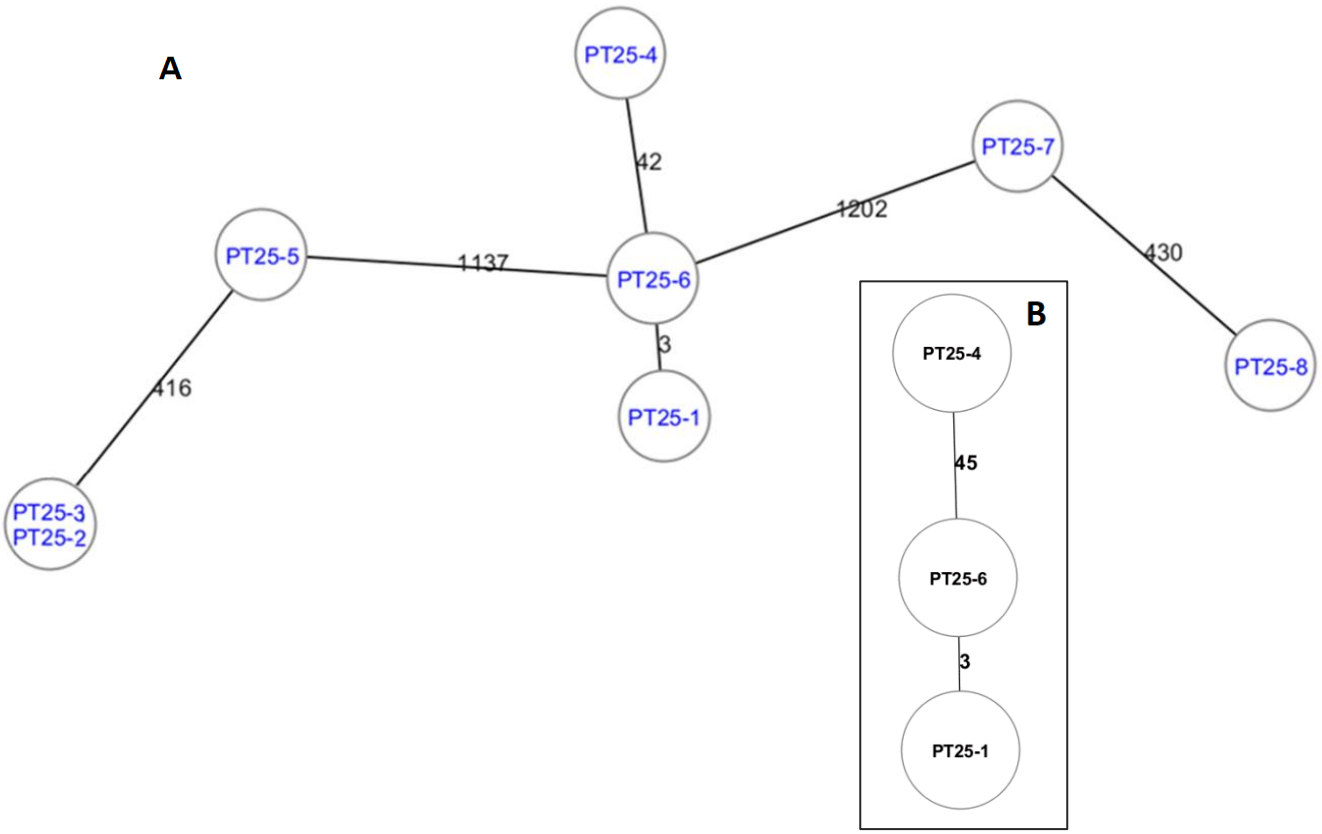


Figure 1. A) Minimum spanning tree created in SeqSphere+ showing a core genome MLST analysis on the eight strains used in PT 25. The sequence data was produced by the EURL and the core genome scheme used was the Oxford v1 comprising 1,343 gene targets. The text in the spheres represents the sample names and the numbers on the lines represent numbers of allele differences. The total number of genes used in the analysis, after discarding genes not present in all assemblies, was 1,247. The line length is not proportional to the number of allele differences. B) The same analysis as in A) but only performed on the three samples clustering together. The separation of the samples from the other samples maximises the resolution of the cluster. For this subset, the number of gene targets compared was 1,286.

In Figure 1, the two strains at the far left (PT25-2 and PT25-3) are the two samples containing DNA from the same strain and they should produce inseparable subtyping results, regardless of technique. The middle cluster containing three strains (PT25-1, PT25-4 and PT25-6) was designed to assess the participants' ability to interpret if isolates are part of an outbreak or not.

Preparation of DNA samples

Strains were cultivated on horse blood agar and the DNA was extracted from colonies using the EZ1 Advanced robot with the DNA tissue kit (Qiagen). The concentration of the extracted DNA was measured using Qubit 2.0 and the DNA HS-kit (Thermo Fisher Scientific). Ten DNA-extractions from the same agar plate were pooled to create a stock for each sample. The DNA was stabilised by adding DNAstable® Plus (Biomatrica) in a ratio 1:4, which ensured that the DNA was stable in room temperature for the duration of the PT. The final concentration of each sample was > 20 ng/µl.

The DNA was aliquoted (> 40 µl) into 1.5 ml DNAse free screw cap microtubes of which three of each sample were kept for quality control during the PT. The quality control samples were sequenced when the samples were prepared, when the PT was dispatched and at the deadline of the PT (June 10th). No differences were noted between runs regarding the created libraries, sequencing quality or the analysis to be performed (ST determination and cluster analysis).

Scenario for the cluster analysis

A scenario was created that mimicked an actual outbreak investigation. The scenario was also to help the participants to better understand how to interpret the cluster analysis and to answer the questions related to that part of the PT.

The following scenario was incorporated into the PT instructions:

- Several cases of campylobacteriosis have been reported from patients who have consumed raw milk purchased from the same vending machine.
- Upon analysis of the milk, 2 *Campylobacter jejuni* isolates are isolated.
- There are 3 different farms that deliver milk to the vending machine.
- Milk filters are collected from the different farms and 6 *C. jejuni* isolates are obtained altogether from farms A, B and C.
- An investigation is launched to establish molecular epidemiological links to the source of the *C. jejuni* in the sold milk.

Distribution of the proficiency test

The PT 25 samples were distributed from the EURL on 11th of March 2019 together with PT 23 and PT 24. 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. Each participant received a plastic sample container containing the eight microtubes along with filling material (paper) to keep sample tubes stable during transport.

A Micro-T-Log was included in each shipment to record the temperature every second hour during transport. The accompanying instructions recommended the participants to refrigerate (+2-8°C) the samples on arrival until the DNA sequencing was to be performed.

The results had to be reported via a Questback survey by 10th of June 2019. A link to the questionnaire was sent out by e-mail in connection with the distribution of the test from the EURL. A copy of the Questback form was included as an appendix to the instructions accompanying the samples.

Reporting

The participants could choose to analyse the samples using either Sanger sequencing or WGS to determine the sequence types.

Questions to be answered by the participants (Question 2 was optional and only for participants using WGS):

1. Which Sequence Types (STs) do the 8 isolates belong to?

2. Please list the isolate/isolates you think is from the same source as sample PT25-1 and sample PT25-2, respectively. (This question will not be scored)

Participants were instructed to derive the STs (or allele identifiers if no complete ST could be determined) from PubMLST.org to simplify performance assessment.

To answer the second question, participants using WGS needed to perform a more in-depth analysis of the relatedness of the samples to determine which samples could be related to each other. This analysis was in the Questback survey referred to as a "cluster analysis" and it could include a core genome MLST (cgMLST) or a whole genome MLST (wgMLST) or a single nucleotide polymorphism (SNP)-analysis.

In addition to question 1 and 2, several questions regarding technologies/methods used had to be answered. Summaries of some of these questions can be found in Appendix A at the end of this document.

Data submission

Participants using WGS had to upload the following to the EURL via the cloud service Onehub:

- Raw data files from the sequencing instrument (FASTQ-files)
- Genome assemblies (if assembly was performed)
- Any result files used to draw conclusions from the cluster analysis. This includes images of, for instance, minimum spanning trees from cgMLST or trees drawn from SNP-analysis.

Participants were given access to the cloud service Onehub and a separate workspace and login was created for each participant.

Assessing the performance

Performance assessment was performed only on the reported eight sequence types (STs). The maximum score was therefore 8. There was no acceptance limit defined for this PT.

For the optional WGS cluster analysis there were no official thresholds to indicate an epidemiological link and cluster analysis was therefore performed based on relative relationships in the dataset and based on the participants' previous experiences from defined outbreaks using their currently established method. The answers given by the participants was their interpretations of the cluster analysis. Therefore, no acceptance limit was defined for this part of the PT. The EURL evaluated how well the reported answer was supported by the data uploaded to the EURL. From the raw data (FASTQ-files) and assemblies uploaded to the EURL, comparisons were performed to evaluate and quantify the differences of the different sequencing runs. The differences between all reported answers was also presented and discussed at the annual EURL workshop in October.

Results

All participants stated that the package was received without any visible signs of outer damage done to the package and all packages were delivered within two days.

All the participants that reported results were MS-NRLs. The number of participants and which sequencing technologies they used is shown in Table 2 and the results from the sequence typing is shown in Table 3 and 4.

Table 2. The number of participating laboratories in PT 25.

Sequencing method used	Sanger	WGS
No. of participants	5	20
No. of results reported	3	18*

*A nineteenth participant uploaded results one day late and is not included in this report.

Table 3. The sequence types (ST) reported by the participants using Sanger sequencing. The eight samples are named PT25-X where X is the sample number. The correct answer is shown in the first line under the samples. Erroneous results have been highlighted. If no complete ST was determined, the number of correct alleles reported is shown.

	PT25-1	PT25-2	PT25-3	PT25-4	PT25-5	PT25-6	PT25-7	PT25-8	Score
Correct ST ->	257	21	21	257	883	257	1326	45	
Lab ID									
37	257	21	6/7 alleles	257	21	257	45	45	5
57	257	21	21	257	883	257	1326	45	8
59	257	21	21	257	6/7 alleles	257	1326	45	7

Table 4. The sequence types (ST) reported by the participants using WGS. The eight samples are named PT25-X where X is the sample number. The correct answer is shown in the first line under the samples. Erroneous results have been highlighted.

	PT25-1	PT25-2	PT25-3	PT25-4	PT25-5	PT25-6	PT25-7	PT25-8	Score
Correct ST ->	257	21	21	257	883	257	1326	45	
Lab ID									
15	257	21	21	257	883	257	1326	45	8
16	257	21	21	257	883	257	1326	45	8
18	257	21	21	257	883	257	1326	45	8
19	257	21	21	257	883	257	1326	45	8
20*	257	21	21	257	883	257	1326	45	8
22	257	21	21	257	883	257	1326	45	8
23	257	21	21	257	883	257	1326	45	8
24	257	21	21	257	883	257	1326	45	8
27	257	21	21	257	883	257	1326	45	8
35	257	21	21	257	883	257	1326	45	8
39	257	21	21	257	883	257	1326	45	8
41	257	21	21	257	883	257	1326	45	8
51	257	21	21	257	883	257	1326	45	8
53	257	21	21	257	883	257	1326	45	8
54	257	21	21	257	883	257	1326	45	8
56	257	21	21	257	21	257	45	45	6
61	257	21	21	257	257	883	1326	45	6
65	257	21	21	257	883	257	1326	45	8

* Lab ID 20 complemented the analysis with Sanger sequencing

As presented in Table 3, two (Lab ID 37 and 59) out of three participants using Sanger sequencing failed to correctly determine all eight STs. For the WGS-participants, two (Lab ID 56 and Lab ID 61) out of 18 participants failed in determining the correct STs. However, the EURL has analysed the raw data and assemblies uploaded by Lab ID 56 and 61 and the correct STs could be determined from both their assemblies and their raw data. This indicates that at least some of the wrong answers could be due to human errors such as mixing up samples at the reporting stage.

Notice should be taken that Lab ID 20 used the IonTorrent S5 machine and due to errors in sequencing over homopolymeric DNA regions commonly made by that technology, Sanger sequencing had to be performed of the *tkt* genes to complete the sequence types. All other WGS-participants used Illumina-based machines and did not have to complement with Sanger sequencing.

Cluster analysis results

Sixteen participants performed the voluntary cluster analysis part of PT 25. As shown in Figure 1, the interpretation the EURL has made is that PT25-3 should be reported as being from the same source as PT25-2. All participants answered that PT25-3 were from the same source as PT25-2, thus supporting the interpretation.

In the cluster involving the three samples: PT25-1, PT25-4 and PT25-6, the EURL has made the interpretation that only sample PT25-6 should be reported as being from the same source as sample PT25-1. Sample PT25-4 aligns closely with the other two samples, regardless of method used. For instance, when applying core genome MLST on the data files submitted by the participants, the number of allele differences were somewhere close to 40 and in concordance with the analysis by the EURL represented in Fig 1. Since the samples were supposed to be part of an ongoing outbreak, 40 allele differences should indicate that PT25-4 is not part of the investigated outbreak. Twelve out of sixteen participants answered that only PT25-6 was from the same source as PT25-1, thus supporting this interpretation.

The four participants answering that also PT25-4 was from the same source were Lab IDs 24, 27, 61 and 65. Some possible explanations to, or discussions about, their differing answers are stated below:

Lab ID 24 used cgMLST but visualised the results as a dendrogram instead of showing the actual number of allele differences and this could make the interpretation more difficult.

Lab ID 27 was the only participant using a software called Mumi and their result image clearly separates sample PT25-4 from the other two samples. However, without having access to more information about actual differences between samples in their analysis, the EURL cannot evaluate the interpretation made by the participant.

Lab ID 61 used the commercial software SeqSphere+ and the cgMLST scheme included in that software. The included cgMLST-scheme only consists of 637 targets, which is relatively low compared to the scheme used by PubMLST that comprises 1,343 targets. The result image uploaded by Lab ID 61 only separates PT25-1 and PT25-4 with 12 allele differences. The default cluster alert function in SeqSphere+ uses 13 allele differences for *C. jejuni/coli* and has therefore highlighted the three samples to be part of a cluster. However, another participant, Lab ID 18, used the same software but also included the so-called accessory gene target scheme, which comprises 958 additional targets. This elevates the number of targets investigated from 637 to 1,595 and the resolution is therefore increased significantly. Lab ID 18 produced a result in which PT25-1 and PT25-4 are separated by over 41 allele differences, clearly differentiating the sample from the outbreak. This highlights the need to investigate enough targets when performing cg/wgMLST based studies.

The result image from the fourth lab, Lab ID 65, showed 14 single nucleotide polymorphisms and 8 possible recombination events between the two discussed samples. They did however answer that PT25-4 is "perhaps also from the same source as PT25-1".

Summary of proficiency test number 25, 2019

The objective of PT 25 was to assess the performance of subtyping eight strains of *C. jejuni* that were delivered as extracted DNA to the participants. The multi locus sequence type (ST) was to be determined for each strain by the participating laboratories and the maximum score was therefore eight. Twenty-five participants signed up for the PT and 21 submitted results before the deadline.

Of the three participants that used Sanger sequencing, one correctly determined all eight STs. The other two participants correctly determined five and seven STs, respectively.

Of the 18 participants that used WGS to determine the STs, 16 correctly determined all eight STs. However, the EURL has analysed the raw data submitted by the participants and the correct STs could be determined from all participants. This suggests that the erroneous results were due to human errors, for instance, at the reporting stage of the PT.

A fictional outbreak investigation was included in the PT that enabled the WGS-participants to perform a cluster analysis to determine which samples that were from the same source. The cluster analysis was not mandatory, and 16 participants submitted results on this part. The interpretations made by most of the participants are the same as made by the EURL. An alternate interpretation on one of the two clusters in the dataset was made by four of the 16 participants. The reason for the differing interpretations could be explained in some cases (e.g., due to low resolution in the analysis).

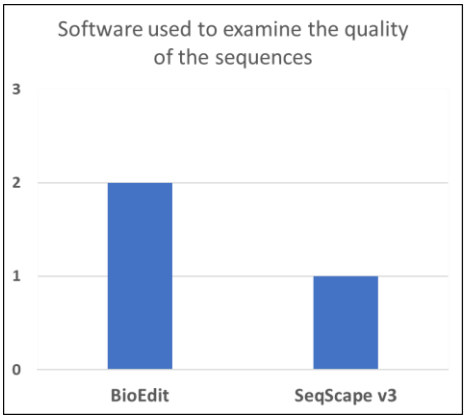
References

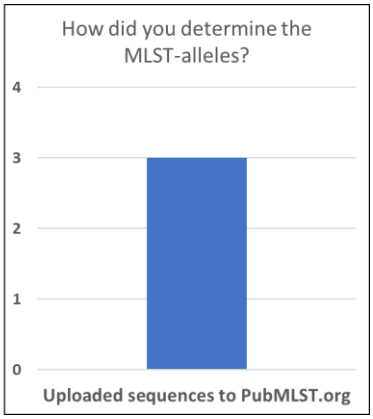
1. Bolger, A.M., M. Lohse, and B. Usadel, *Trimmomatic: a flexible trimmer for Illumina sequence data.* Bioinformatics, 2014. **30**(15): p. 2114-20.
2. Bankevich, A., et al., *SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing.* J Comput Biol, 2012. **19**(5): p. 455-77.
3. Walker, B.J., et al., *Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement.* PLoS One, 2014. **9**(11): p. e112963.
4. Cody, A.J., et al., *Core Genome Multilocus Sequence Typing Scheme for Stable, Comparative Analyses of Campylobacter jejuni and C. coli Human Disease Isolates.* J Clin Microbiol, 2017. **55**(7): p. 2086-2097.

Appendix A

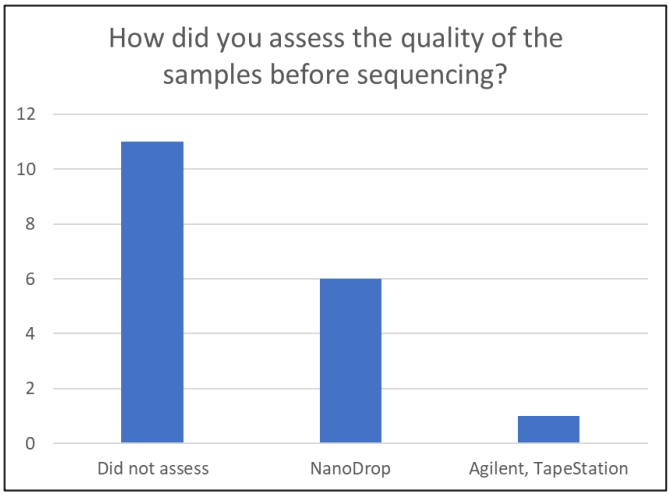
The following diagrams summarises the information provided by the participants regarding technologies and methods used by the participants of PT 25.

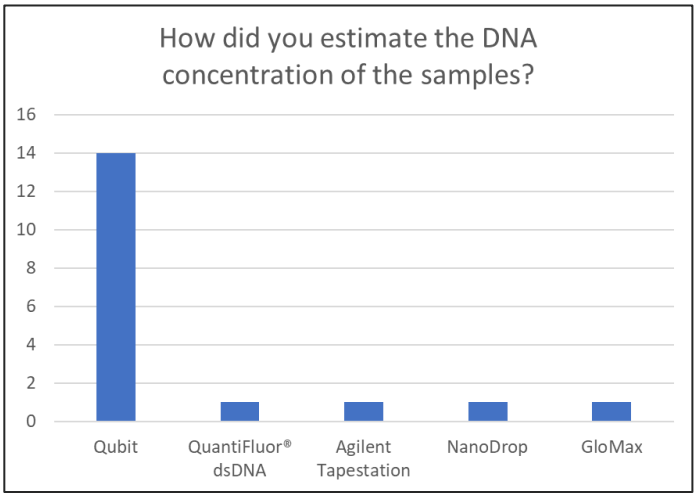
The following two questions were answered by participants using Sanger sequencing.

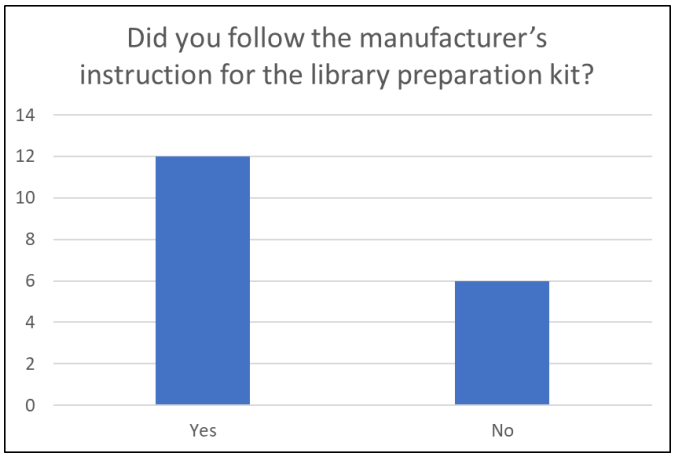




The following questions were answered by participants using WGS.







Changes made to the manufacturer's instructions by participants answering "No" in the above question:

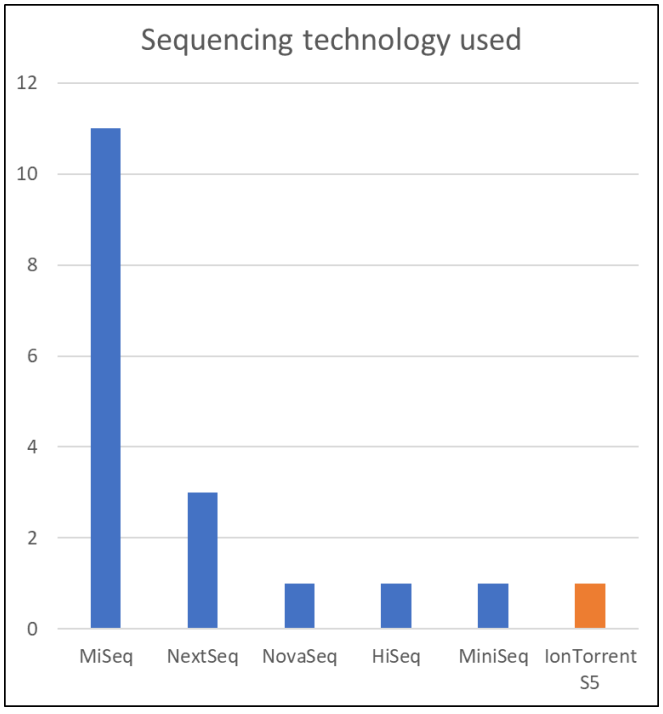
2 participants – Used 50% of volumes

1 participant – Used 40% of volumes

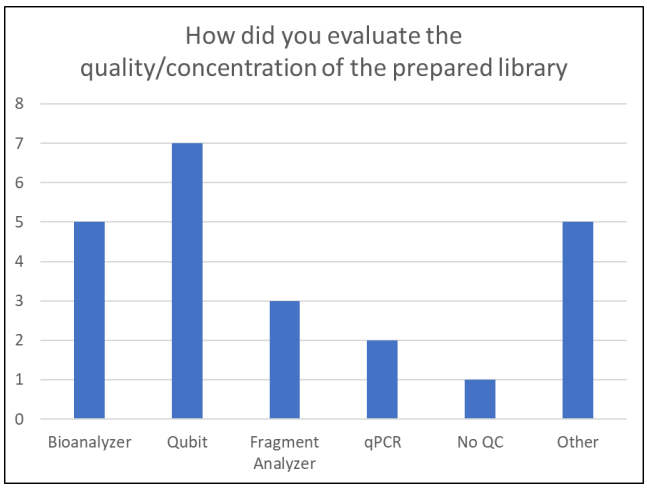
1 participant – Used 20% of volumes

1 participant – Used 13 µl instead of 15 µl of NPM (PCR-master mix in Nextera XT-kit)

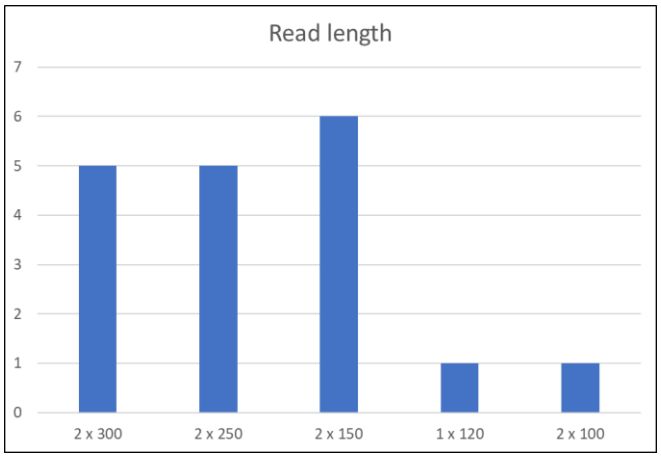
1 participant – Used minor changes to the PCR-program performed when using the Nextera XT-kit



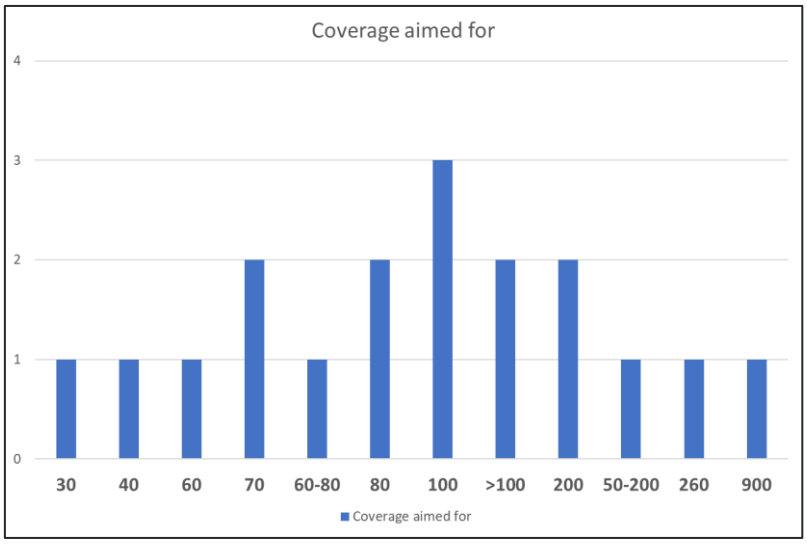
All the Illumina-based technologies are shown in blue colour and IonTorrent-based technologies are shown in orange.



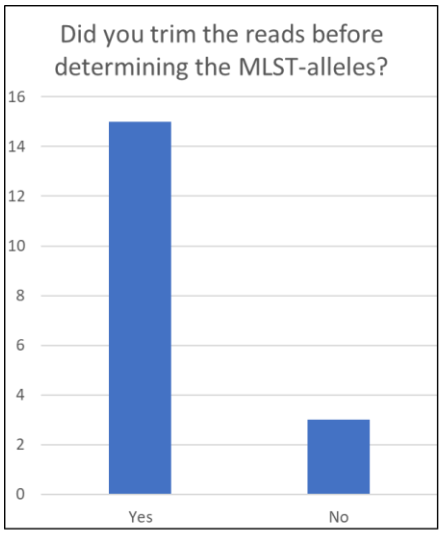
In the above diagram, some participants have used combinations of the methods.

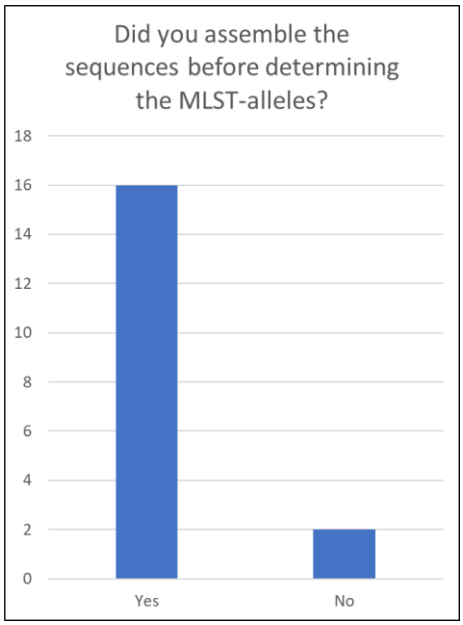


The above diagram summarises the read length used for sequencing. The number "1" or "2" before the read lengths indicates if the run was a single-end run ("1") or a paired-end run ("2").

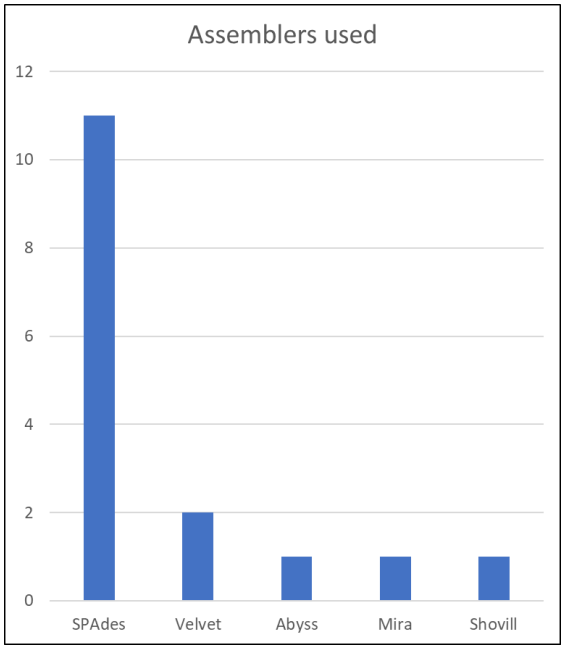


The above diagram summarises the amount of data the participants aimed for when sequencing. The theoretical times coverage of the sample genome yielded by a sequencing run is dependent on the level of multiplexing (i.e., the amount of samples sequenced in the same run) in relation to the kit size (how many bases that can be achieved by the kit and machine). The actual coverage yielded is also dependant on other factors such as cluster density and quality of the run.



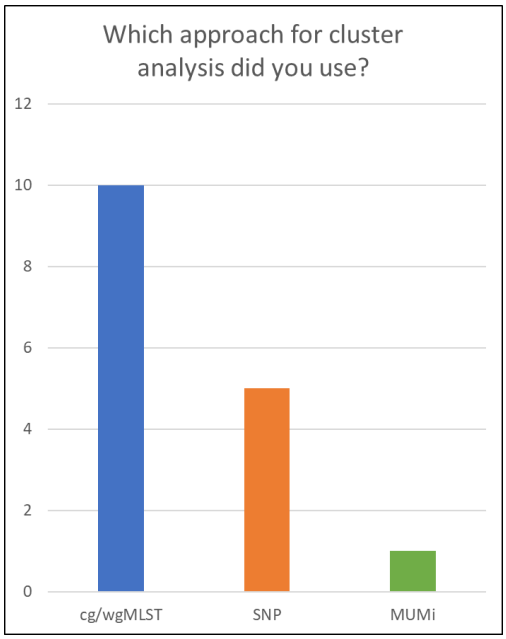


The above diagram summarises whether the reads were trimmed for quality before the MLST-alleles were determined and whether the reads were assembled into draft genomes before the alleles were determined.

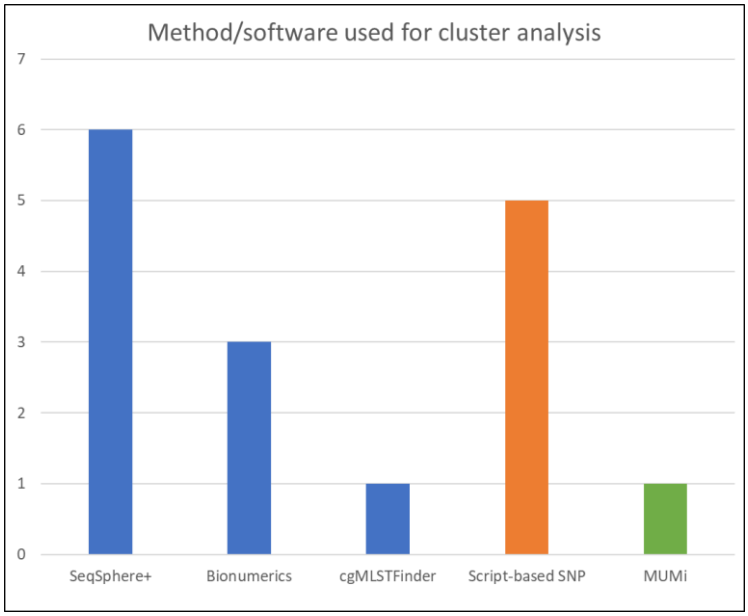


The above diagram summarises the different assembly software used to produce draft genomes from which the MLST-alleles could be determined from.

The following questions were answered by participants that performed the optional cluster analysis part of PT 25.



In the above diagram, the number of participants using cg/wgMLST, SNP-analysis or MUMi, respectively, are shown. The method of using MUMi has not been evaluated by the EURL and possibly qualifies as a SNP-method.



In the above diagram, the MLST-based methods are shown in blue, the SNP-based methods are shown in orange and the MUMi based approach is shown in green.