
SWEDISH
VETERINARY
AGENCY

Proficiency test number 38

WGS and Cluster Analysis of *Campylobacter*

Bo Segerman and Ásgeir Ástvaldsson

EURL-*Campylobacter* workshop

Uppsala, October 22-23rd, 2024

SWEDISH VETERINARY AGENCY

EURL *Campylobacter*

Co-funded by the European Union

Co-funded by the European Union. Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or the European Health and Digital Executive Agency (HaDEA). Neither the European Union nor HaDEA can be held responsible for them.

Objective

Assess quality of WGS data and accuracy of sequence analysis of *Campylobacter* from participating laboratories

Purpose

To help laboratories implement and evaluate their capacity of WGS and sequence analysis of Campylobacter

Participation

- **24 NRLs in 17 EU Member states and in Iceland, Switzerland and the UK registered for PT38**

- **All NRLs reported results to the EURL**
  - One NRL only reported results on the sequence analysis part

Outline

Divided in two parts

- **Part 1**: Library preparations and sequencing of two DNA samples
- Samples for part 1 were distributed together with PT36, 11th of March 2024

- **Part 2**: QC, species identification, MLST and cluster analysis of a dataset containing 18 raw data sequence samples
- Dataset for part 2 was available for download from OneDrive on the same day as the PTs were distributed

Part 1

- **Samples should be processed according to standard laboratory procedures**
  - DNA reconstitution > DNA quantification and QC > Library preparations > Sequencing

TABLE 1. Identity of the two DNA samples distributed to the NRLs in proficiency test No. 38, 2024.

Sample ID	Species	Sampling year	Sequence type (ST)	GC content (%)	Genome size
PT38-1	*Campylobacter jejuni*	2016	464	30.29	1.8 Mb
PT38-2	*Campylobacter coli*	2017	4709	31.18	1.8 Mb + 0.1 Mb plasmid

- **Reference genomes available from PT28**

Part 2

- **Dataset containing 18 raw data sequence samples from *Campylobacter***
- **Three samples in-silico modified**
  - **PT38-10**: contaminated with 10% *P. aeruginosa* reads
  - **PT38-15**: modified so only 13% bases were Q30+ > too little high-quality data to be useful
  - **PT38-19**: contaminated with 40% *S. enterica*

Perform QC, identify species, designate MLST and perform cluster analysis

TABLE 2. Identity of the 18 raw-data samples distributed to the NRLs in proficiency test No. 38, 2024.

Sample ID	Species	Location	Sampling time	Sequencer	Amount of data (Mbases)	% Q30+	Q30+ base coverage (X)	% reads with adapters ≥12 bp
PT38-3	*C. jejuni*	Farm A	Oct., 2020	NovaSeq (151+151)	455	90	250	19
PT38-4	*C. jejuni*	Farm B	Oct., 2020	MiSeq (251+251)	212	93	120	0.1
PT38-5	*C. jejuni*	Farm C	Sep., 2019	NovaSeq (151+151)	417	86	220	73
PT38-6	*C. jejuni*	Farm C	Sep., 2021	MiSeq (76+76)	149	95	91	0.003
PT38-7	*C. jejuni*	Farm D	Jun, 2021	NextSeq 500 (151+151)	366	89	223	0.14
PT38-8	*C. jejuni*	Farm E	Nov., 2020	NovaSeq (151+151)	438	93	268	29
PT38-9	*C. jejuni*	Farm D	Aug., 2018	MiSeq (76+76)	499	95	290	0.002
PT38-10	*C. jejuni*	Farm F	Jul., 2020	MiSeq (76+76)	198	90	121	0.008
PT38-11	*C. coli*	Farm G	Sep., 2021	NovaSeq (151+151)	400	90	230	36
PT38-12	*C. jejuni*	Farm C	Sep., 2021	NovaSeq (151+151)	423	88	256	73
PT38-13	*C. coli*	Farm G	Sep., 2021	NovaSeq (151+151)	485	90	280	56
PT38-14	*C. jejuni*	Farm C	Sep., 2019	NovaSeq (151+151)	413	85	253	66
PT38-15	*C. jejuni*	Farm H	Aug., 2017	NovaSeq (151+151)	88	13	7	
PT38-16	*C. jejuni*	Farm I	Sep., 2020	MiSeq (251+251)	250	93	152	0.1
PT38-17	*C. jejuni*	Farm J	Oct., 2020	MiSeq (251+251)	192	93	117	0.1
PT38-18	*C. jejuni*	Farm K	Oct., 2020	NovaSeq (151+151)	242	92	220	73
PT38-19	*C. jejuni*	Farm L	Aug., 2019	MiSeq (76+76)	124	95	91	0.03
PT38-20	*C. jejuni*	Farm C	Sep., 2021	NovaSeq (151+151)	450	90	275	45

Reporting

- **Deadline:** 15th of May 2024
- Through a Questback questionnaire

Requested data uploaded to a personal OneDrive folder

- Part 1:

  Raw sequence files (i.e. fastq files)

- Part 2:

  Assembly files (FASTA), if part of analysis

  Tree used to draw conclusions (e.g. phylogenetic tree or mst)

  Raw clustering data used to create trees (e.g. distance matric or alignment)

Part 1 - Assessment of sequence quality

Cut-off values defined for six different criteria to assess the sequence quality of two DNA samples

TABLE 3. Overview of the criteria and cut-off values used for assessment of sequence quality in proficiency test No. 38 (2024).

Criteria	Cut-off value for satisfactory performance
Total amount of data	>30X or 80X depending on library preparation kit (80X for Nextera XT)
Q30+	>70 %, 75 % or 80 % depending on read length (300, 250, 150-100 bp)
Contamination	<5 % from non-target species
Reference coverage	>98 % of reference genome[a]
GC deviation	<4 % deviation from reference genomes
Assembly targets	>95 % of targets found

[a]The maximum amount of data used for the assessment was 80X coverage for NRLs using Nextera XT and 30X coverage for NRLs using other library preparation kits.

Part 1 - Results

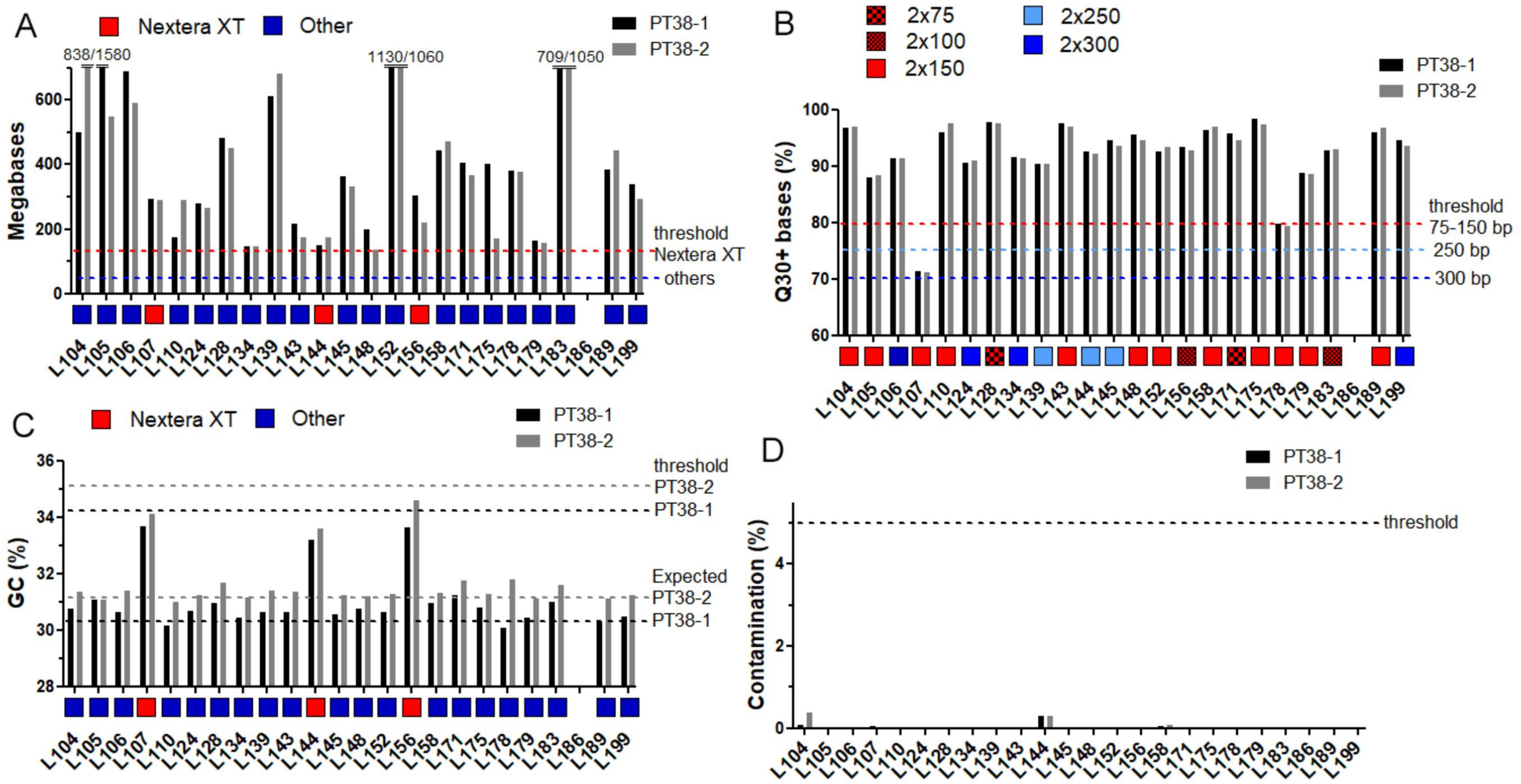


FIGURE 1. **A:** Total amount of data submitted by the NRLs. Thresholds were set at 30X for unbiased library preps (Blue) and 80x for Nextera XT preps (Red). **B:** Percentage of Q30+ bases in the WGS data submitted by the NRLs. Thresholds are set based on the read length used. **C:** Deviation of the GC content in the reads from the expected GC content (the GC content of the reference genome). **D:** Contamination levels estimated by the Kraken2 software using the 30 GB standard database.

Part 1 - Results

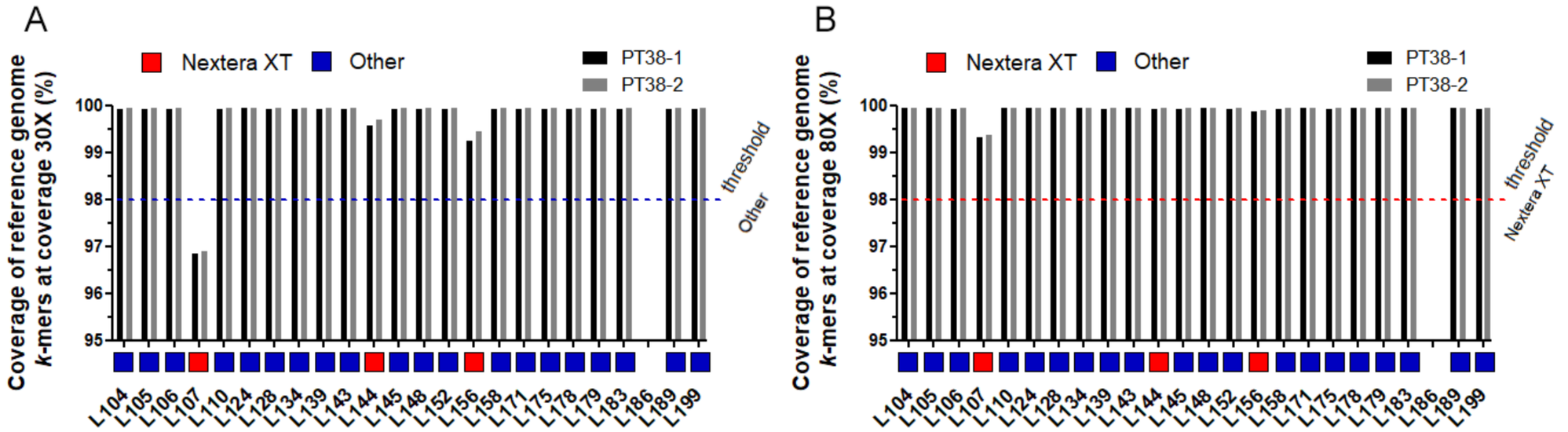


FIGURE 2. **A:** Coverage of the reference genome (k-mers) in the read data at 30X sequencing depth. The threshold applies for non-Nextera XT library kits. **B:** Coverage of the reference genome (k-mers) in the read data at 80X sequencing depth. The threshold applies for Nextera XT library kits.

Part 1 - Results

- 21 NRLs fulfilled the criteria for satisfactory performance
- 2 NRLs scored below the criteria for satisfactory performance

TABLE 5. Overview of assessment of the sequence quality of each NRL in proficiency test No. 38 (2024). The number indicate number of samples out of two reaching the criteria cut-offs.

Lab ID	Amount of data	Q30+	Contamination	Reference coverage	GC deviation	Assembly targets	Overall evaluation sequence quality
L104	2/2	2/2	2/2	2/2	2/2	2/2	Satisfactory
L105	2/2	2/2	2/2	2/2	2/2	2/2	Satisfactory
L106	2/2	2/2	2/2	2/2	2/2	2/2	Satisfactory
L107	2/2	0/2	2/2	2/2	2/2	2/2	Needs improvement
L110	2/2	2/2	2/2	2/2	2/2	2/2	Satisfactory
L124	2/2	2/2	2/2	2/2	2/2	2/2	Satisfactory
L128	2/2	2/2	2/2	2/2	2/2	2/2	Satisfactory
L134	2/2	2/2	2/2	2/2	2/2	2/2	Satisfactory
L139	2/2	2/2	2/2	2/2	2/2	2/2	Satisfactory
L143	2/2	2/2	2/2	2/2	2/2	2/2	Satisfactory
L144	2/2	2/2	2/2	2/2	2/2	2/2	Satisfactory
L145	2/2	2/2	2/2	2/2	2/2	2/2	Satisfactory
L148	2/2	2/2	2/2	2/2	2/2	2/2	Satisfactory
L152	2/2	2/2	2/2	2/2	2/2	2/2	Satisfactory
L156	2/2	2/2	2/2	2/2	2/2	2/2	Satisfactory
L158	2/2	2/2	2/2	2/2	2/2	2/2	Satisfactory
L171	2/2	2/2	2/2	2/2	2/2	2/2	Satisfactory
L175	2/2	2/2	2/2	2/2	2/2	2/2	Satisfactory
L178	2/2	0/2	2/2	2/2	2/2	2/2	Needs improvement
L179	2/2	2/2	2/2	2/2	2/2	2/2	Satisfactory
L183	2/2	2/2	2/2	2/2	2/2	2/2	Satisfactory
L189	2/2	2/2	2/2	2/2	2/2	2/2	Satisfactory
L199	2/2	2/2	2/2	2/2	2/2	2/2	Satisfactory

Assessment of sequence quality (part 2)

Cut-off values defined for four different criteria, which were all assessed separately

TABLE 4. Overview of the criteria and cut-off values used for assessment of sequence analysis in proficiency test No. 38 (2024).

Criteria	Cut-off value for satisfactory performance
Evaluation of sequence quality	Identify and exclude (or 'clean up') sample PT38-10, PT38-15 and PT38-19
Species identification	All samples analysed[a] should be correctly species identified
MLST determination	All samples analysed[a] should be designated with correct ST
Cluster detection	Cluster A (or AB), C and D should be identified

[a]Samples PT38-10, PT38-15, PT38-19 excluded from the assessment.

Part 2 – Results

Evaluation of Sequence Quality

Three samples were in-silico modified.

Participants were expected to exclude these samples from analysis

- 18 of 24 NRLs excluded contaminated samples
- All NRLs excluded the low-quality samples
- 7 of 24 NRLs excluded additional samples

TABLE 6. Overview of results from the participants' evaluation of sequence quality in Part 2 of proficiency test No. 38 (2024).

Lab ID	Excluded PT38-10 (10 % Pseudomonas)	Excluded PT38-19 (40 % Salmonella)	Excluded PT38-15 (low QC score)	Number of other exclusions
L104	Yes[a]	Yes[a]	Yes	0
L105	No	Yes	Yes	2
L106	Yes	Yes	Yes	0
L107	Yes	Yes	Yes	0
L110	No	No	Yes	0
L124	Yes	Yes	Yes	1
L128	Yes	Yes	Yes	0
L134	No	Yes	Yes	0
L139	No	No	Yes	0
L143	Yes	Yes	Yes	3
L144	Yes	Yes	Yes	0
L145	Yes	Yes	Yes	0
L148	Yes[a]	Yes	Yes	2
L152	No	Yes	Yes	0
L156	Yes	Yes	Yes	1
L158	Yes	Yes	Yes	0
L171	Yes	Yes	Yes	0
L175	Yes	Yes	Yes	0
L178	Yes	Yes	Yes	0
L179	Yes	Yes	Yes	0
L183	Yes	Yes	Yes	1
L186	Yes[a]	Yes	Yes	0
L189	No	Yes	Yes	0
L199	Yes	Yes	Yes	8

[a]Sample excluded from the analysis in the supplementary data, but exclusion was not reported in Questback.

Part 2 – Results

Species identification and MLST

Participants were expected to identify the species and determine the ST

22 of 24 identified correct species in all samples

21 of 24 determined the ST correctly

TABLE 7. Overview of results from the MLST determination in Part 2 of proficiency test No. 38 (2024).

Lab ID	PT38-3	PT38-4	PT38-5	PT38-6	PT38-7	PT38-8	PT38-9	PT38-10	PT38-11	PT38-12	PT38-13	PT38-14	PT38-15	PT38-16	PT38-17	PT38-18	PT38-19	PT38-20
L104	257	257	257	148	257	257	257		854	148	854	257		257	257	257		148
L105	257	257	257	148	257	257	257		854	148	854	257		257	257	257		148
L106	257	257	257	148	257	257	257		854	148	854	257		257	257	257		148
L107	257	257	257	148	257	257	257		854	148	854	257		ND[a]	257	257		148
L110	257	257	257	148	257	257	257		854	148	854	257		257	257	257		148
L124	257	257	257	148	257	257	257		854	148	854	257		257	257	257		148
L128	257	257	257	21	257	257	257		828	21	828	257		257	257	257		21
L134	257	257	257	148	257	257	257		854	148	854	257		257	257	257		148
L139	257	257	257	148	257	257	257		854	148	854	257		257	257	257		148
L143	257	257	EXCL	148	257	257	257		854	EXCL	854	EXCL		257	257	257		148
L144	257	257	257	148	257	257	257		854	148	854	257		257	257	257		148
L145	257	257	257	148	257	257	257		854	148	854	257		257	257	257		148
L148	257	257	EXCL	148	257	257	257		854	148	854	257		EXCL	257	257		148
L152	257	257	257	148	257	257	257		854	148	854	257		257	257	257		148
L156	257	257	257	148	257	257	257		854	148	854	257		257	257	257		148
L158	257	257	257	148	257	257	257		854	148	854	257		257	257	257		148
L171	257	257	257	148	257	257	257		854	148	854	257		257	257	257		148
L175	257	257	257	148	257	257	257		854	148	854	257		257	257	257		148
L178	257	257	257	148	257	257	257		854	148	854	257		257	257	257		148
L179	257	257	257	148	257	257	257		854	148	854	257		257	257	257		148
L183	257	257	257	148	257	257	257		ND[b]	148	ND[b]	257		257	257	257		EXCL
L186	257	257	257	148	257	257	257		854	148	854	257		257	257	257		148
L189	257	257	257	148	257	257	257		854	148	854	257		257	257	257		148[c]
L199	257	257	257	148	257	257	257		854	148	854	257		257	257	257		148

[a] Missing alleles.

[b] All seven alleles were correctly identified, but ST was not determined.

[c] Wrongly reported in Questback, but correctly determined in the supplementary uploaded data.

Part 2 – Results

Cluster detection

	PT38-3	PT38-4	PT38-8	PT38-16	PT38-18	PT38-17	PT38-9	*PT38-14*	*PT38-5*	PT38-7	PT38-6	*PT38-12*	*PT38-20*	*PT38-11*	*PT38-13*	PT38-15	PT38-10	PT38-19	cut-off	schema	software
L158	A	A	A	A	A	A	A	A	A	A	C	C	C	D	D	EXCL	EXCL	EXCL	NA	PubMLST v1	cgMLSTFinder
L107	A	A	A	A	A	A	A	A	A	A	C	C	C	D	D	EXCL	EXCL	EXCL	NA	NA	snippy
L199	A	A	A	A	A	A	EXCL	EXCL	EXCL	EXCL	NO	EXCL	EXCL	EXCL	EXCL	EXCL	EXCL	EXCL	7 AD	PubMLST v1	ChewBBACA
L106	A	A	A	A	A	A	A	A	A	NO	C	C	C	D	D	EXCL	EXCL	EXCL	13 AD	Ridom core	Ridom SeqSphere
L183	A	A	A	A	A	A	A	A	A	NO	C	C	EXCL	D	D	EXCL	EXCL	EXCL	13 AD	Ridom core	Ridom SeqSphere
L152	A	A	A	A	A	A	A	A	A	NO	C	C	C	D	D	EXCL	C	EXCL	13 AD	Ridom core	Ridom SeqSphere
L189	A	A	A	A	A	A	A	A	A	NO	C	C	C	D	D	EXCL	NO	EXCL	7-10 AD	Ridom core	Ridom SeqSphere
L186	A	A	A	A	A	A	A	A	A	NO	C	C	C	D	D	EXCL	EXCL	EXCL	13 AD	Ridom core	Ridom SeqSphere
L143	A	A	A	A	A	A	A	EXCL	EXCL	NO	C	EXCL	C	D	D	EXCL	EXCL	EXCL	6 SNPs /6 AD	Innuendo core	ChewBBACA/in house SNP
L145	A	A	A	A	A	A	A	B	B	NO	C	C	C	D	D	EXCL	EXCL	EXCL	0.5% AD	Innuendo WG	ChewBBACA
L178	A	A	A	A	A	A	A	B	B	NO	C	C	C	D	D	EXCL	EXCL	EXCL	10 AD	PubMLST v1	ChewBBACA
L139	A	A	A	A	A	A	A	B	B	NO	C	C	C	D	D	EXCL	NO	NO	13 AD	PubMLST v1	Ridom SeqSphere
L156	A	A	A	EXCL	A	A	A	B	B	NO	C	C	C	D	D	EXCL	EXCL	EXCL	10 SNPs	NA	SnapperDB
L105	A	A	A	A	A	A	A	B	B	NO	C	C	C	EXCL	EXCL	EXCL	NO	EXCL	10 AD	Innuendo core	ChewBBACA
L148	A	A	A	EXCL	A	A	A	NO	EXCL	NO	C	C	C	D	D	EXCL	EXCL	EXCL	10 AD	PubMLST v1	Ridom SeqSphere
L124	A	A	A	EXCL	A	A	NO	B	B	NO	C	C	C	D	D	EXCL	EXCL	EXCL	5 AD	PubMLST v1	Ridom SeqSphere
L128	A	A	A	A	A	A	NO	B	B	NO	C	C	C	D	D	EXCL	EXCL	EXCL	5 AD	PubMLST v1	BioNumerics
L179	A	A	A	A	A	A	NO	B	B	NO	C	C	C	D	D	EXCL	EXCL	EXCL	10 AD	PubMLST v1	Ridom SeqSphere
L175	A	A	A	A	A	A	NO	B	B	NO	C	C	C	D	D	EXCL	EXCL	EXCL	10 AD	PubMLST v1	Ridom SeqSphere
L110	A	A	A	A	A	A	NO	B	B	NO	C	C	C	D	D	EXCL	C	NO	NA	NA	Samtools SNP Phylogeny
L134	A	A	A	A	A	A	NO	B	B	NO	C	C	C	D	D	EXCL	NO	EXCL	14 AD	PubMLST v1	ChewBBACA
L104	A	A	A	A	A	A	NO	B	B	NO	C	C	C	D	D	EXCL	EXCL	EXCL	10 SNPs / 31 AD	Innuendo WG	CSI Phylogeny
L171	A	A	A	A	A	A	NO	B	B	NO	C	C	C	D	D	EXCL	EXCL	EXCL	14 AD	Ridom core+Acc	Ridom SeqSphere
L144	A	A	A	A	A	NO	NO	NO	NO	NO	C	NO	C	D	D	EXCL	EXCL	EXCL	5 AD	PubMLST v1	in house

FIGURE 3. Depiction of the different clusters identified by the NRLs

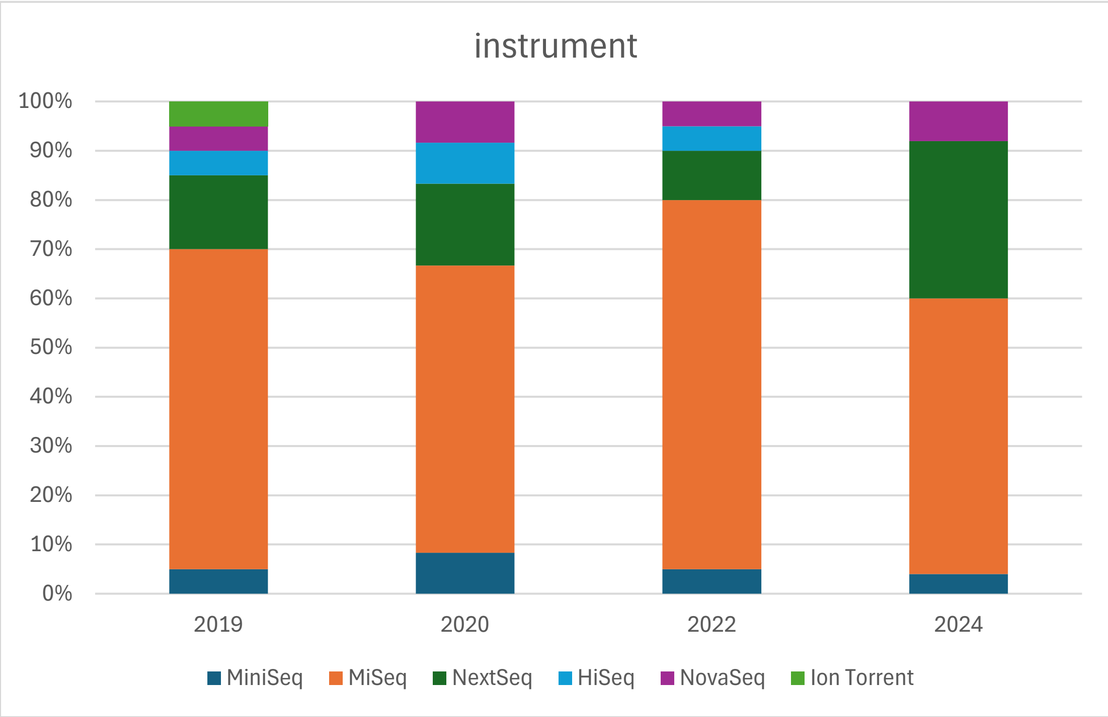
Part 2 - Results

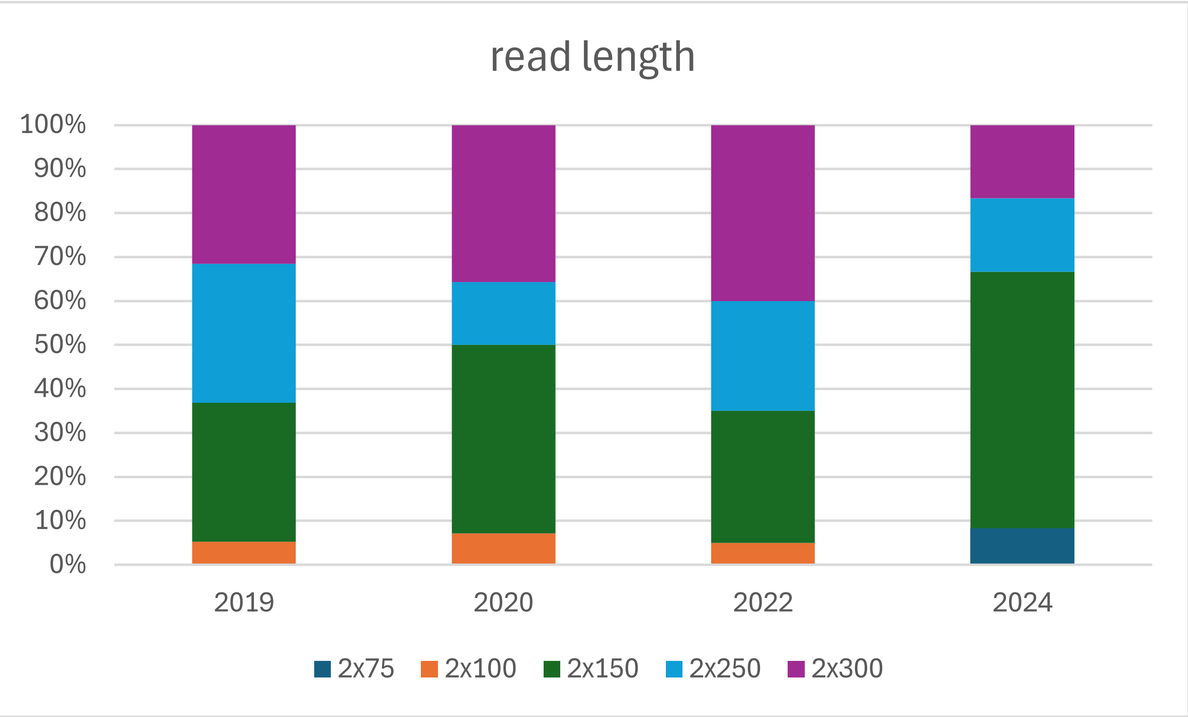
- 14 NRLs fulfilled the criteria for satisfactory performance
- 10 NRLs scored below the criteria for satisfactory performance

TABLE 8. Overview of assessment of the sequence analysis of each NRL in proficiency test No. 38 (2024). The number indicate number of samples out of analysed samples reaching the criteria cut-offs.

Lab ID	Evaluation of sequence quality	Species identification	MLST determination	Cluster detection	Overall evaluation sequence quality
L104	3/3	15/15	15/15	AB, C and D	Satisfactory
L105	2/3	15/15	15/15	AB and C	Needs improvement
L106	3/3	15/15	15/15	A, C and D	Satisfactory
L107	3/3	15/15	14/15	A, C and D	Needs improvement
L110	1/3	15/15	15/15	AB, C and D	Needs improvement
L124	3/3	15/15	15/15	AB, C and D	Satisfactory
L128	3/3	Not performed	10/15	AB, C and D	Needs improvement
L134	2/3	15/15	15/15	AB, C and D	Needs improvement
L139	1/3	15/15	15/15	AB, C and D	Needs improvement
L143	3/3	15/15	12/12	A, C and D	Satisfactory
L144	3/3	15/15	15/15	A, C and D	Satisfactory
L145	3/3	15/15	15/15	AB, C and D	Satisfactory
L148	3/3	13/13	13/13	A, C and D	Satisfactory
L152	2/3	15/15	15/15	A, C and D	Needs improvement
L156	3/3	15/15	15/15	AB, C and D	Satisfactory
L158	3/3	15/15	15/15	A, C and D	Satisfactory
L171	3/3	15/15	15/15	AB, C and D	Satisfactory
L175	3/3	15/15	15/15	AB, C and D	Satisfactory
L178	3/3	15/15	15/15	AB, C and D	Satisfactory
L179	3/3	15/15	15/15	AB, C and D	Satisfactory
L183	3/3	15/15	12/14	A, C and D	Needs improvement
L186	3/3	15/15	15/15	A, C and D	Satisfactory
L189	2/3	14/15	15/15	A, C and D	Needs improvement
L199	3/3	15/15	15/15	A	Needs improvement

Trends

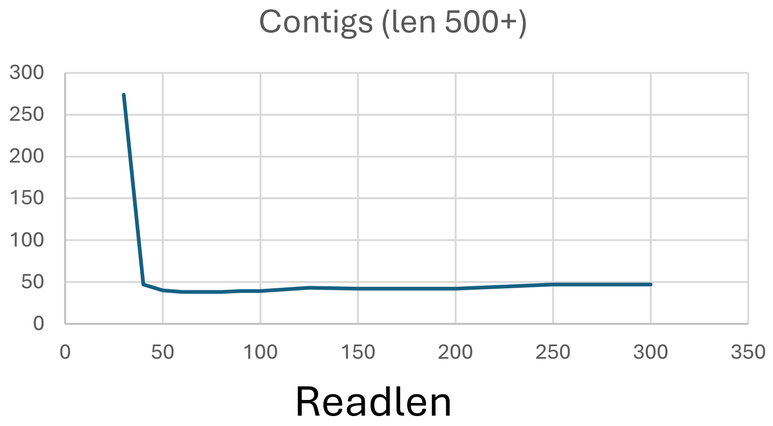


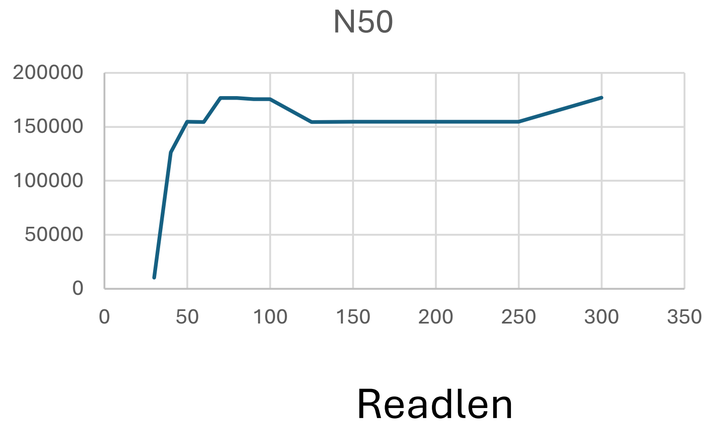


PRESS RELEASE

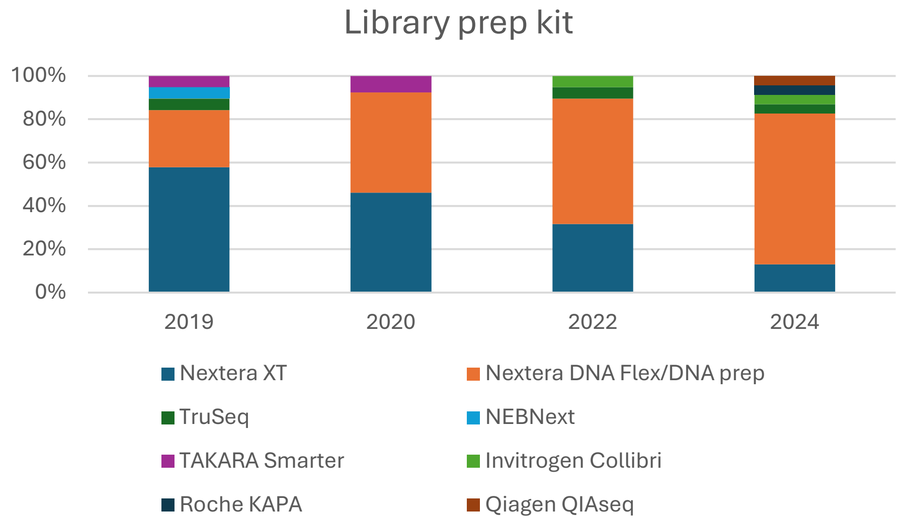
Illumina introduces the MiSeq i100 Series: its simplest, fastest benchtop sequencers

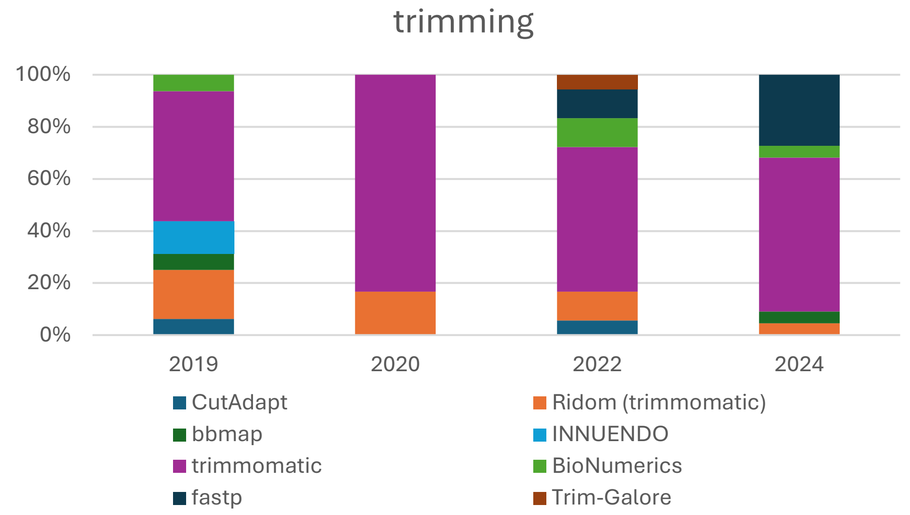
Oct 9, 2024

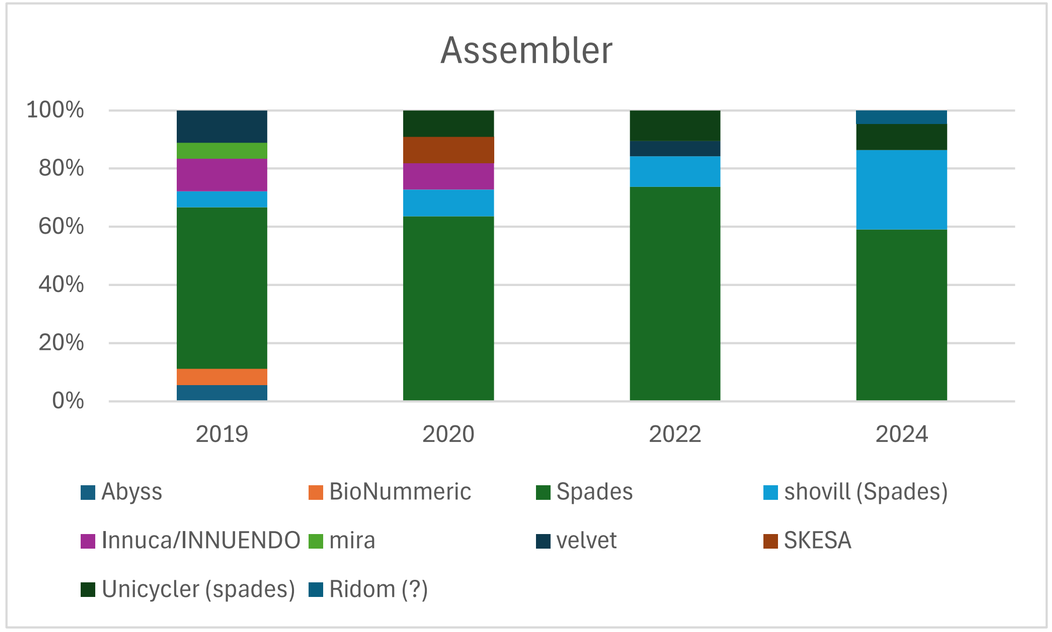


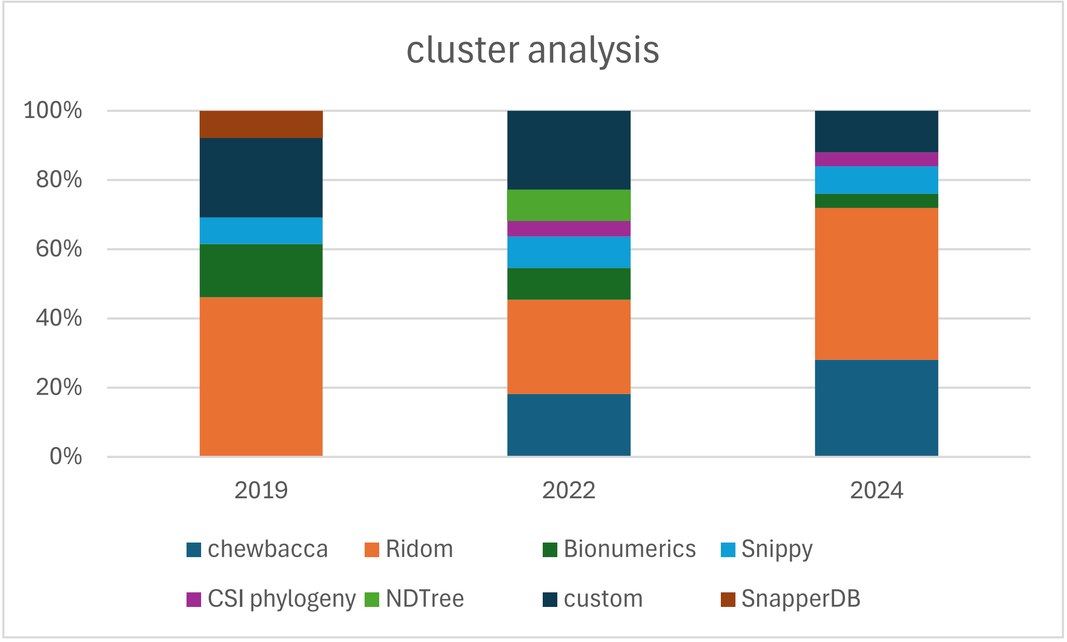


Trends


Library prep kit
100%
80%
60%
40%
20%
0%
2019
2020
2022
2024
Nextera XT
Nextera DNA Flex/DNA prep
TruSeq
NEBNext
TAKARA Smarter
Invitrogen Collibri
Roche KAPA
Qiagen QIAseq


trimming
100%
80%
60%
40%
20%
0%
2019
2020
2022
2024
CutAdapt
Ridom (trimmomatic)
bbmap
INNUENDO
trimmomatic
BioNumerics
fastp
Trim-Galore


Assembler
100%
80%
60%
40%
20%
0%
2019
2020
2022
2024
Abyss
BioNummeric
Spades
shovill (Spades)
Innuca/INNUENDO
mira
velvet
SKESA
Unicycler (spades)
Ridom (?)


cluster analysis
100%
80%
60%
40%
20%
0%
2019
2022
2024
chewbacca
Ridom
Bionumerics
Snippy
CSI phylogeny
NDTree
custom
SnapperDB

Next WGS PT

- **EURL-*Campylobacter* will offer a WGS PT 2026**
- **Similar as PT38**
  - WGS part – DNA samples for sequencing
  - Sequence analysis part – Dataset for sequence and cluster analysis

Thank you!

bo.segerman@sva.se
asgeir.astvaldsson@sva.se

www.sva.se


SWEDISH
VETERINARY
AGENCY