Proficiency test number 25
Subtyping of *Campylobacter jejuni* using MLST or WGS

Joakim Skarin

Outline

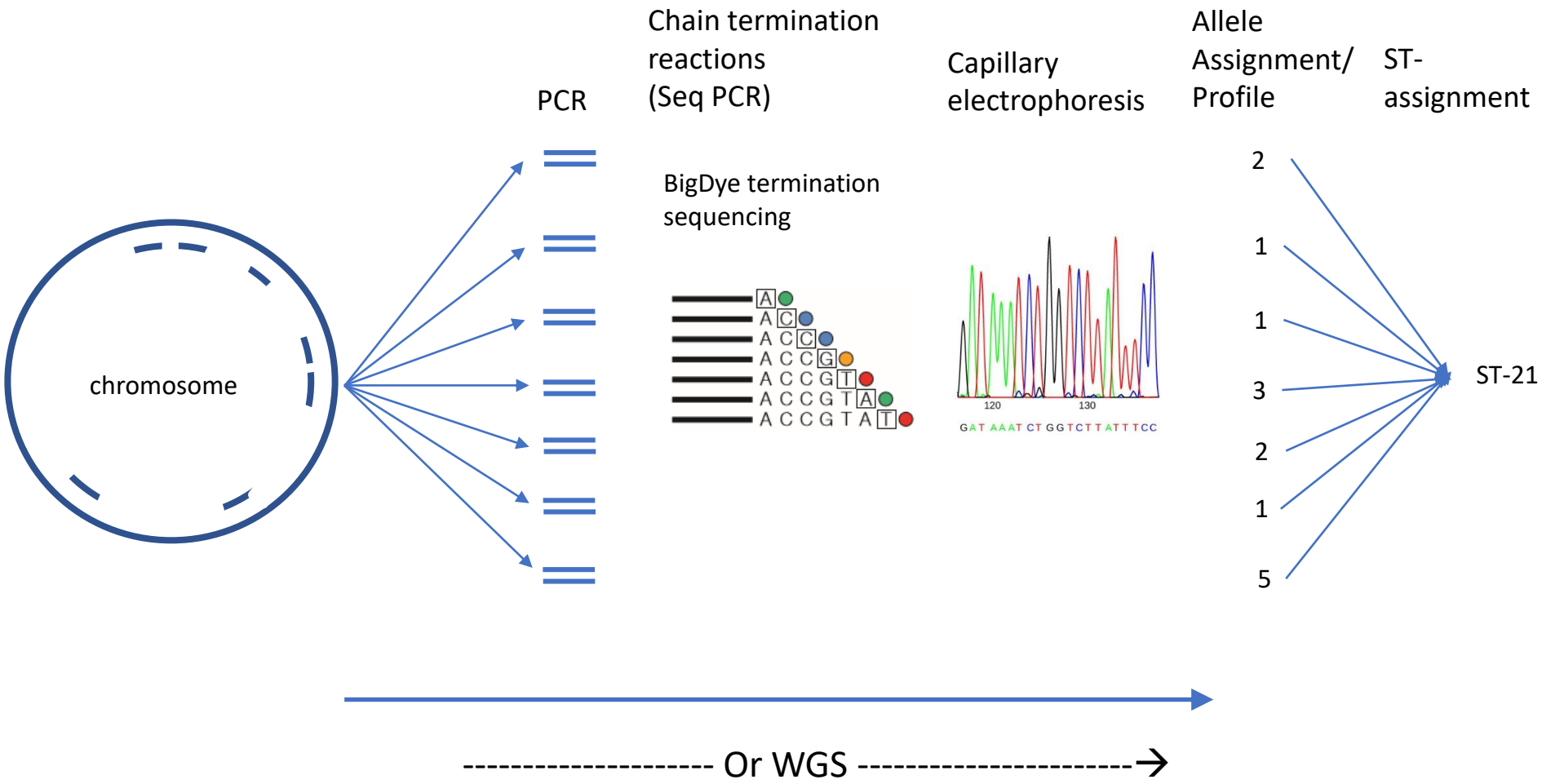
- Definitions and MLST
- Introduction and scenario
- Contents and procedure PT25
- Results, Sanger
- A typical WGS workflow
- Questback answers, WGS
- Results, WGS
- Cluster analysis results and discussion

Definitions

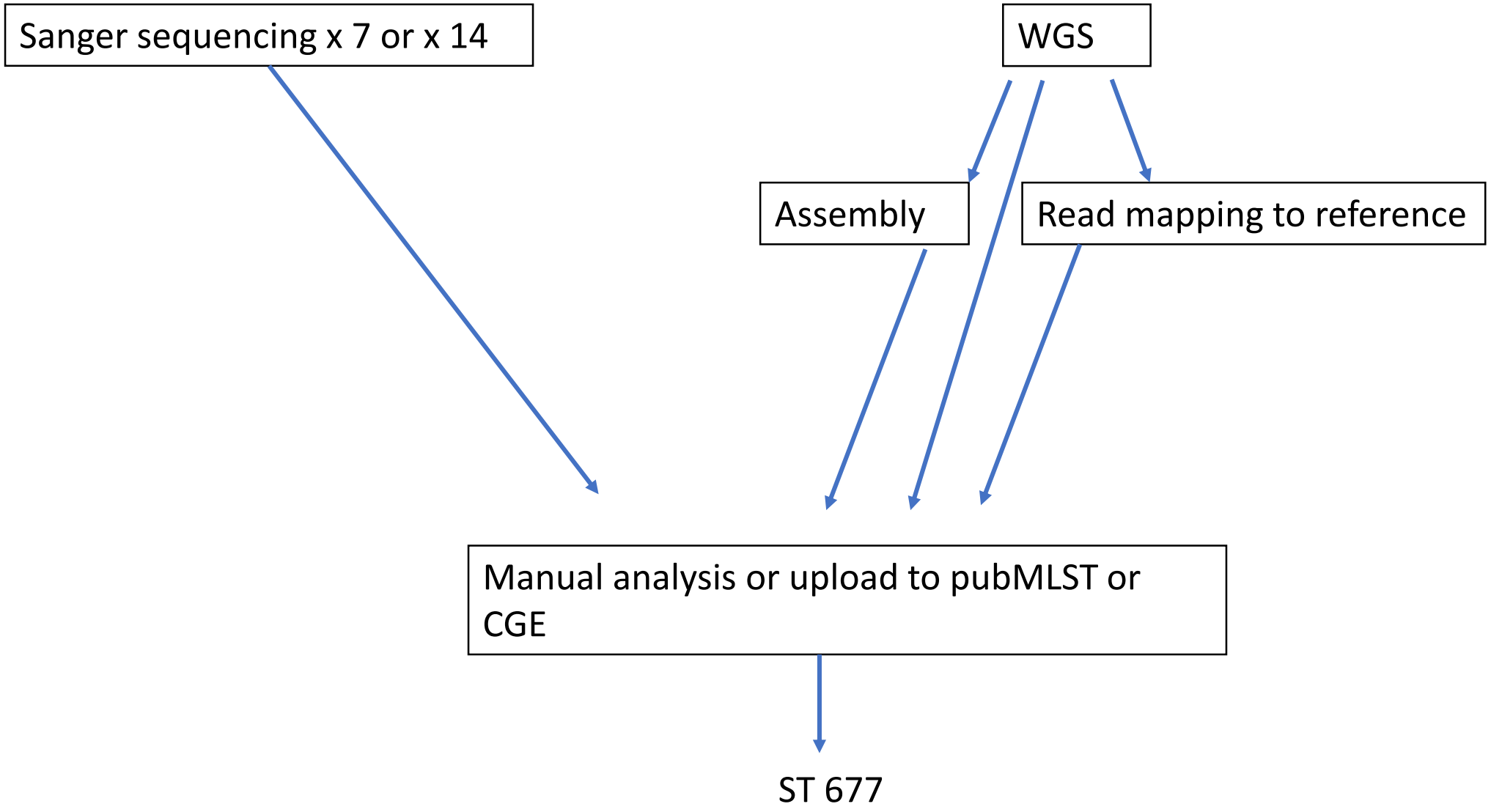
- **Loci:** Loci are regions of the genome that are identified by similarity to a known sequence. It can be nucleotide or peptide sequences. They are often complete coding sequences (genes) but may represent gene fragments (such as used in MLST).

- **Allele:** Alleles are instances of loci. Every unique sequence (either DNA or peptide depending on the locus), is defined as a new allele and they are given an allele identifier in a sequence definition database. Allele identifiers are allocated in the order of discovery.

MLST


chromosome
PCR
Chain termination reactions (Seq PCR)
BigDye termination sequencing
Capillary electrophoresis
Allele Assignment/ Profile
ST-assignment
2
1
1
3
2
1
5
ST-21
Or WGS

From DNA to Sequence Type (ST)



Introduction and scenario – PT25

Pilot-PT

Eight samples of extracted DNA from *Campylobacter jejuni*.

The aim of the PT was to determine the Sequence Type (ST) of the samples using either Sanger-based Multi Locus Sequence Typing (MLST) or Whole Genome Sequencing (WGS).

Participants using WGS could optionally also perform a cluster analysis (e.g., cgMLST or SNP-typing) to determine which samples cluster together.

Scenario:

- 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.

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

1. Which Sequence Types (STs) do the 8 isolates belong to?
2. Does any of the isolates from the milk filters sampled at the farms match any of the isolates in the milk from the vending machine (PT25-1 and PT25-2)? (This question will not be scored)

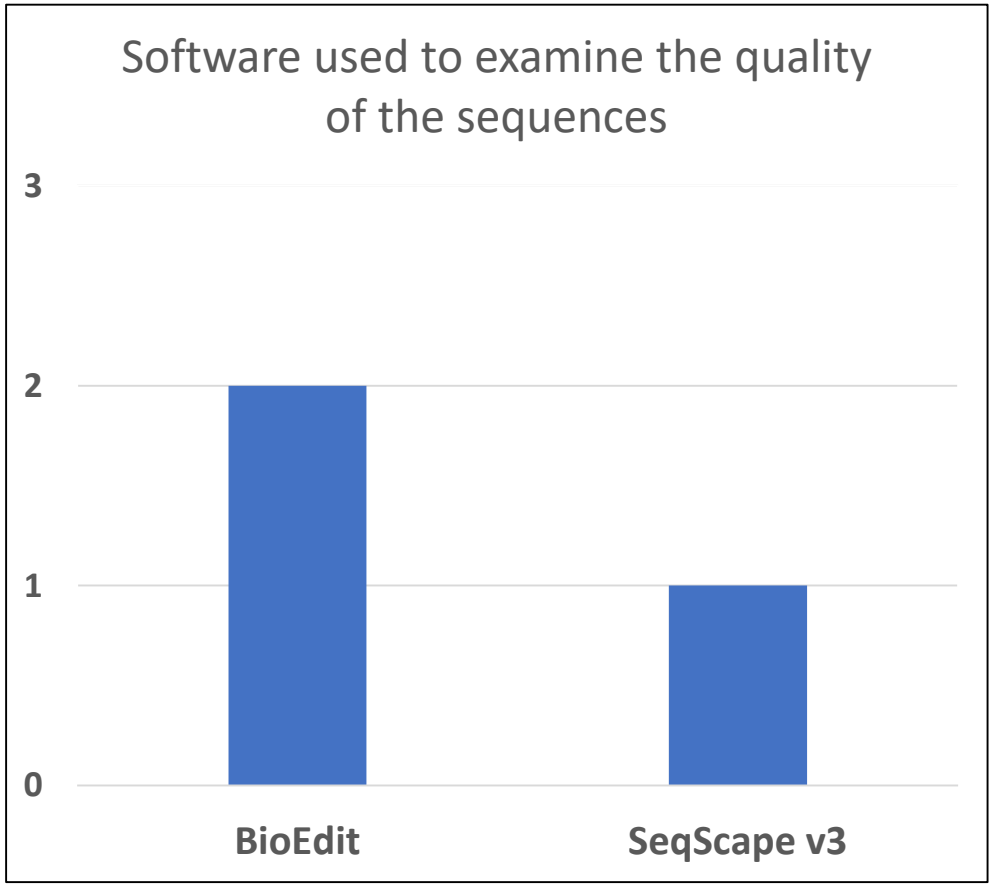
Contents and procedure PT25

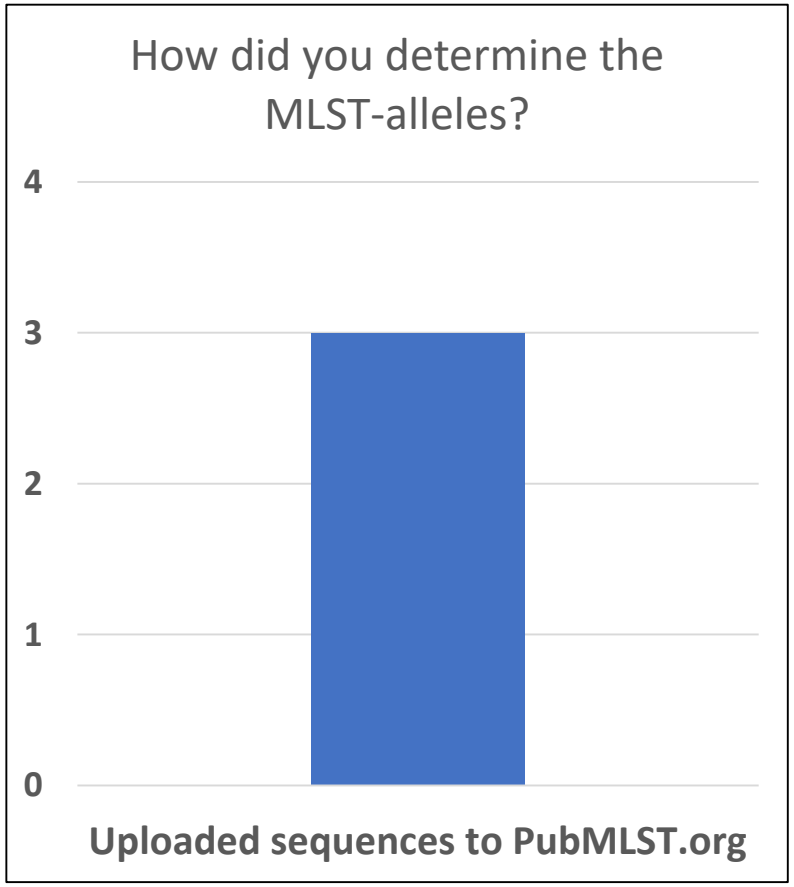
- DNA extracted using Qiagen EZ1 Advanced.
- Concentrations measured using Qubit 2.0
- One large stock per isolate (> 1 ml) of > 20 ng/µl.
- Stabilized using DNAstable® Plus (Biomatrica)
- PT25 sent together with PT23 and PT24
- QC – sequenced after extraction, when leaving SVA and at deadline of PT
- Deadline June 10th to answer Questback survey on methods and results
  - library kits, sequencing machine, read-lengths etc.
  - DNA-concentrations measured, assembly or not, SNP or MLST etc.
  - Results – ST and clustering results
- Onehub – cloud service, one workspace per participating lab, no cost per uploader
  - FASTQ
  - FASTA (if assembly)
  - Supporting images (trees)

Participation in PT25

- 25 signed up for the PT (5 Sanger and 20 WGS)
- 19 WGS labs submitted results
  - 1 lab was 1 day late - included here but not in PT-report
- 3 Sanger labs submitted results

Procedures used by participants, Sanger



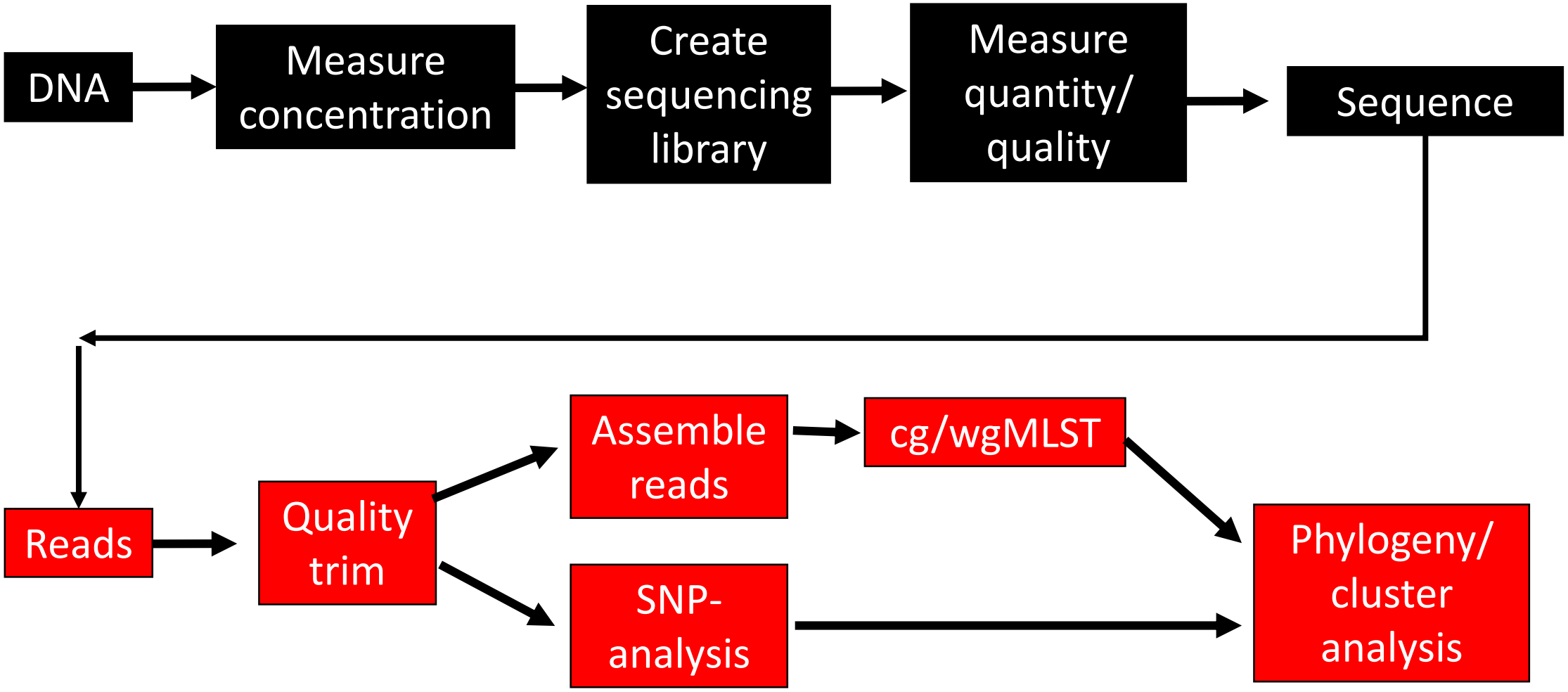


Results, Sanger

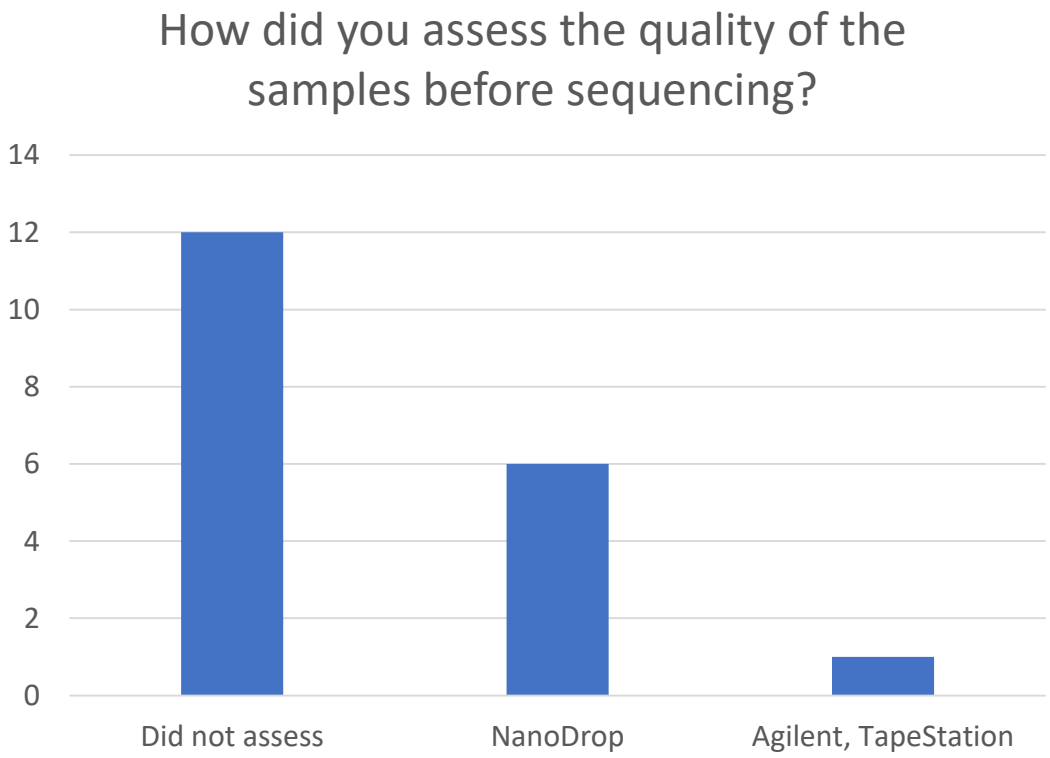
Lab ID	PT25-1	PT25-2	PT25-3	PT25-4	PT25-5	PT25-6	PT25-7	PT25-8
Correct STs	257	21	21	257	883	257	1326	45
57	257	21	21	257	883	257	1326	45
59	257	21	21	257	6/7 correct	257	1326	45
37	257	21	6/7 correct	257	21	257	45	45

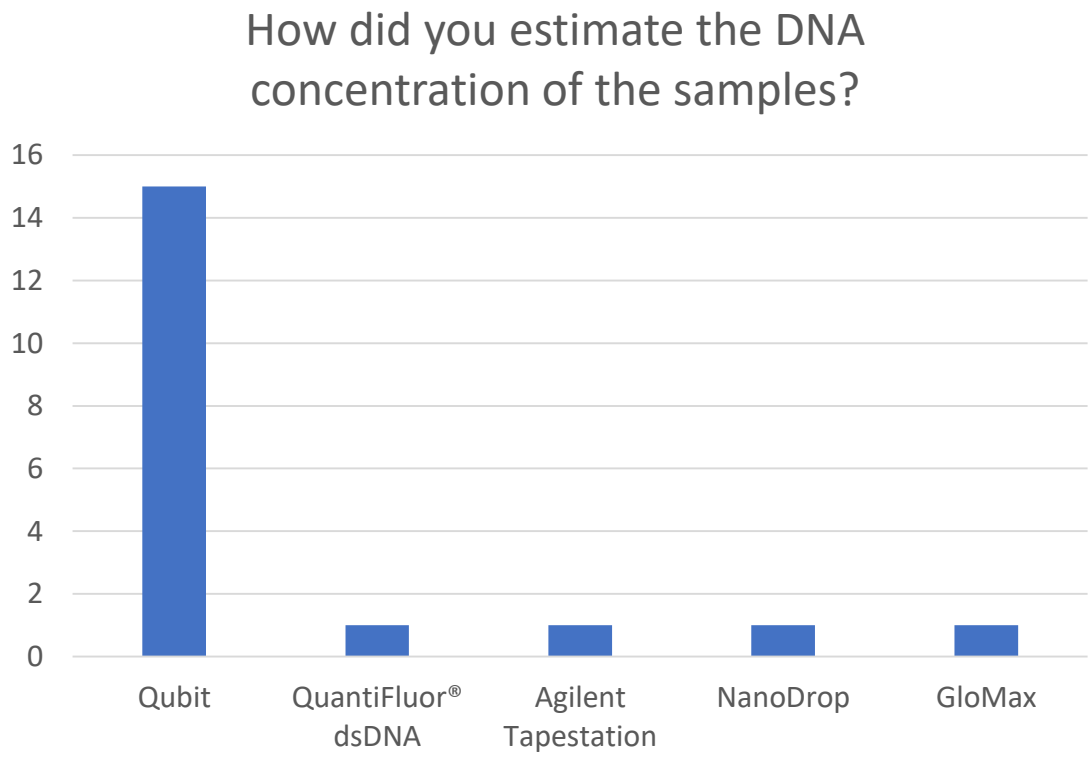
1 / 3 labs – 8 STs determined correctly

A typical WGS workflow



Procedures used by participants, WGS

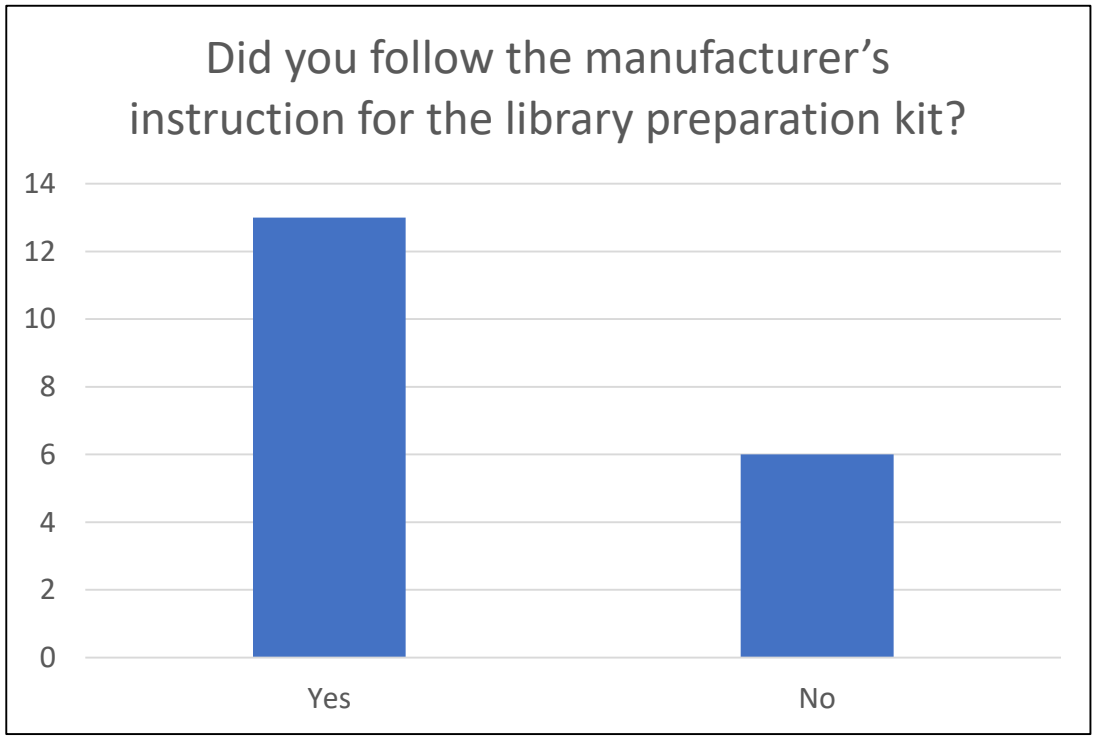




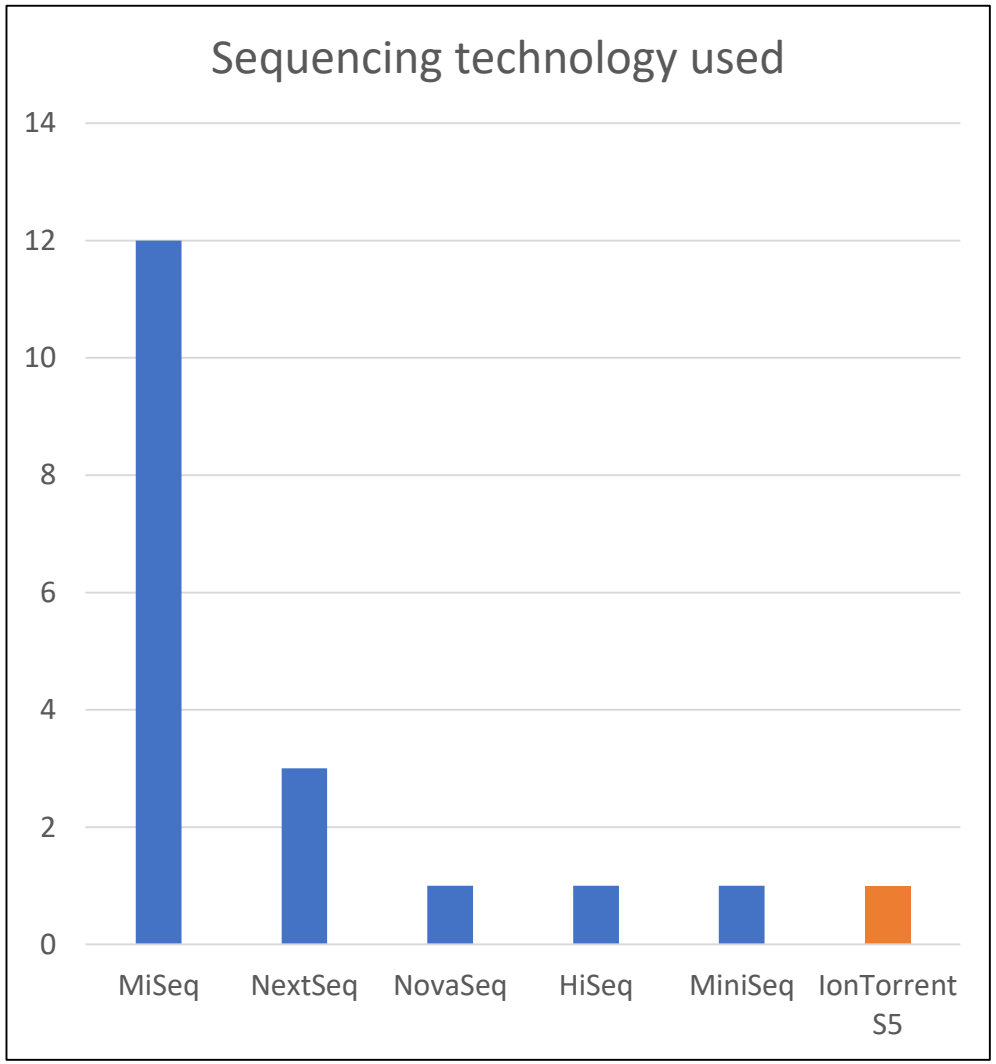
DNA Concentration Measurements

Lab ID	PT25-1	PT25-2	PT25-3	PT25-4	PT25-5	PT25-6	PT25-7	PT25-8
15	14,8	61	37,1	17,8	39,2	25,9	77,8	19
16	29,37	28,97	28,84	24,61	38,16	30,52	34,69	35,89
18	28,6	28,4	27,6	24,1	28,6	26,6	30,7	28,2
19	28,7	25,7	25,8	24,7	28,7	29,7	30,3	28,3
20	39,2	45	39,4	33,6	37,8	41,6	41	43,4
22	35	32	32	31	33	31	39	32
23	36,6	35,6	34,2	23,4	34	33,7	39,1	34,2
24	28,7	25,9	28,2	22,8	26	26,2	30,1	28,8
27	22	26,1	25	23,1	28,5	26,7	29,2	32,5
35	37	38	42	23,5	40	39	41	30
39	24,5	25,8	24,4	20,5	25,3	23,5	26,5	23,7
41	7,94	8,1	8,12	6,82	7,82	7,66	7,28	7,52
51	34,6	35,5	40,5	35,3	41,7	45,3	37,7	37,8
53	87,3	73,6	71,8	74,1	66,6	82,4	69,8	65,6
54	20	20	20	20	20	20	20	20
56	84,3	71	68,5	71,4	64,4	79,7	67,3	63,4
61	37,4	36,3	36,8	37,1	36,8	36,4	39	37
65	39,2	36,4	37,7	32,3	37,6	40,1	40,1	38,3

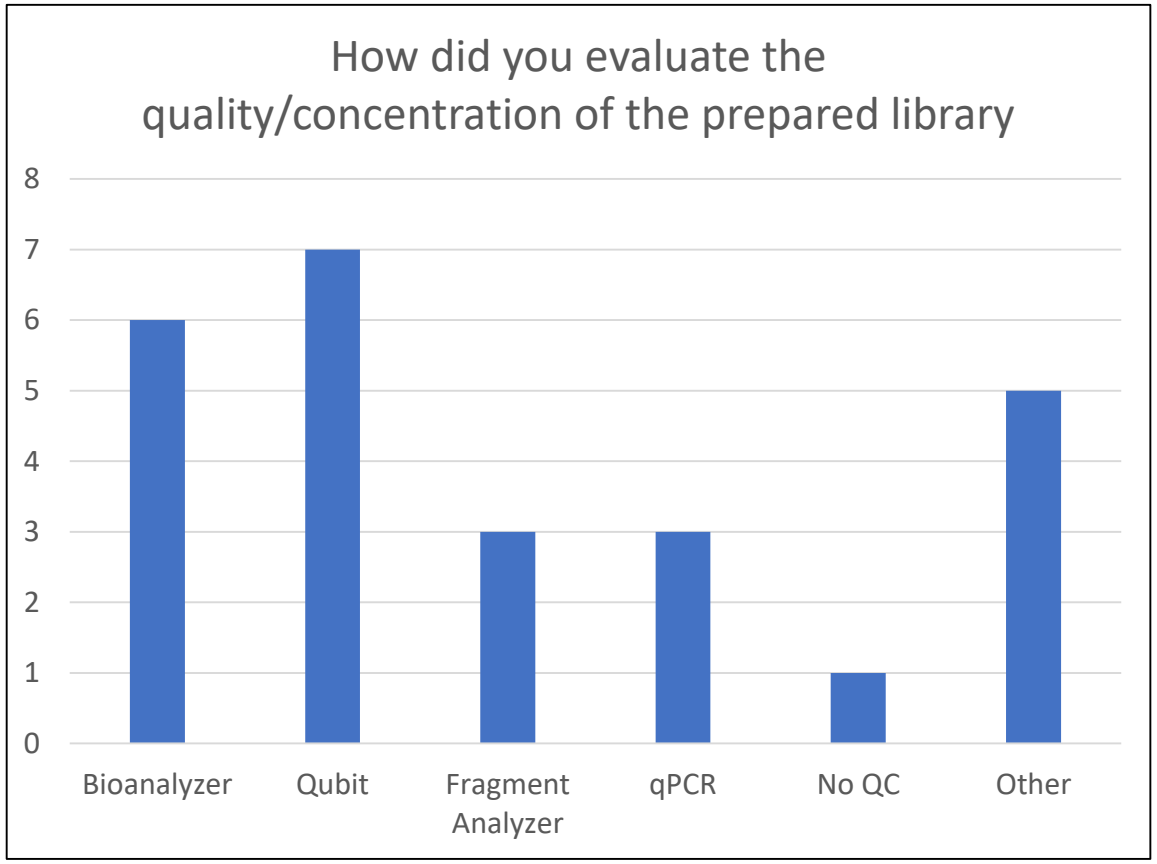
Procedures used by participants, WGS

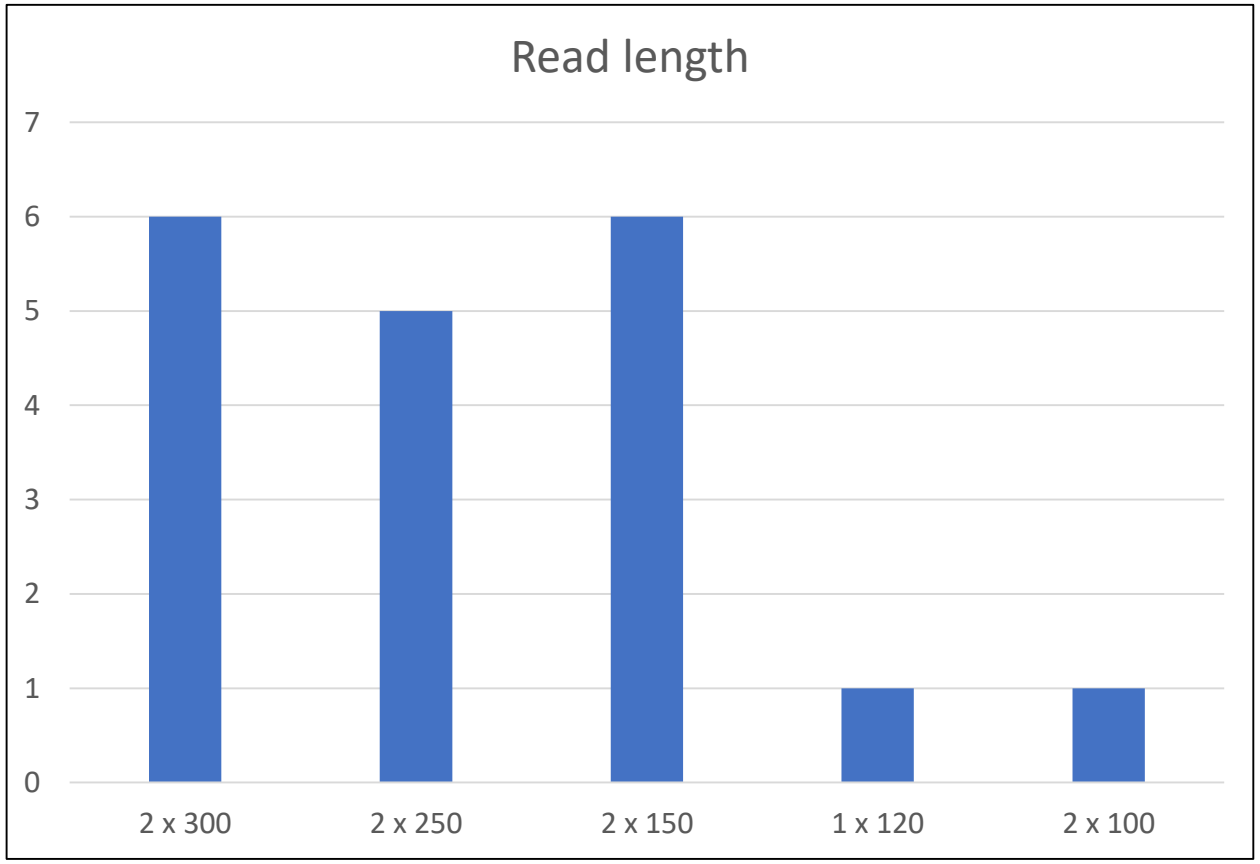


- 2 labs – 50% of volumes
- 1 lab – 40% of volumes
- 1 lab – 20% of volumes
- 1 lab – 13 instead of 15 µl of NPM
- 1 lab – changes to PCR-program

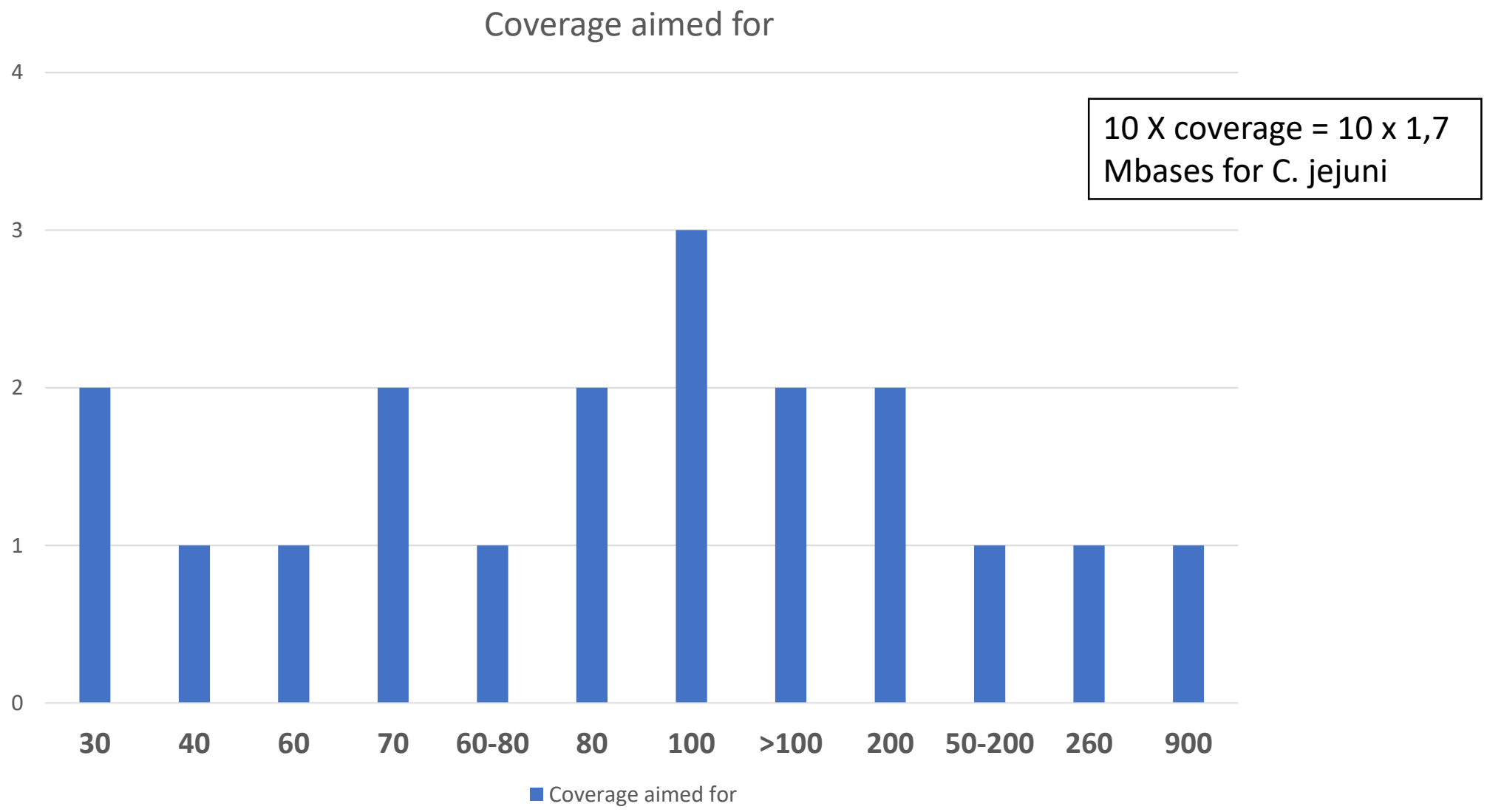


Procedures used by participants, WGS

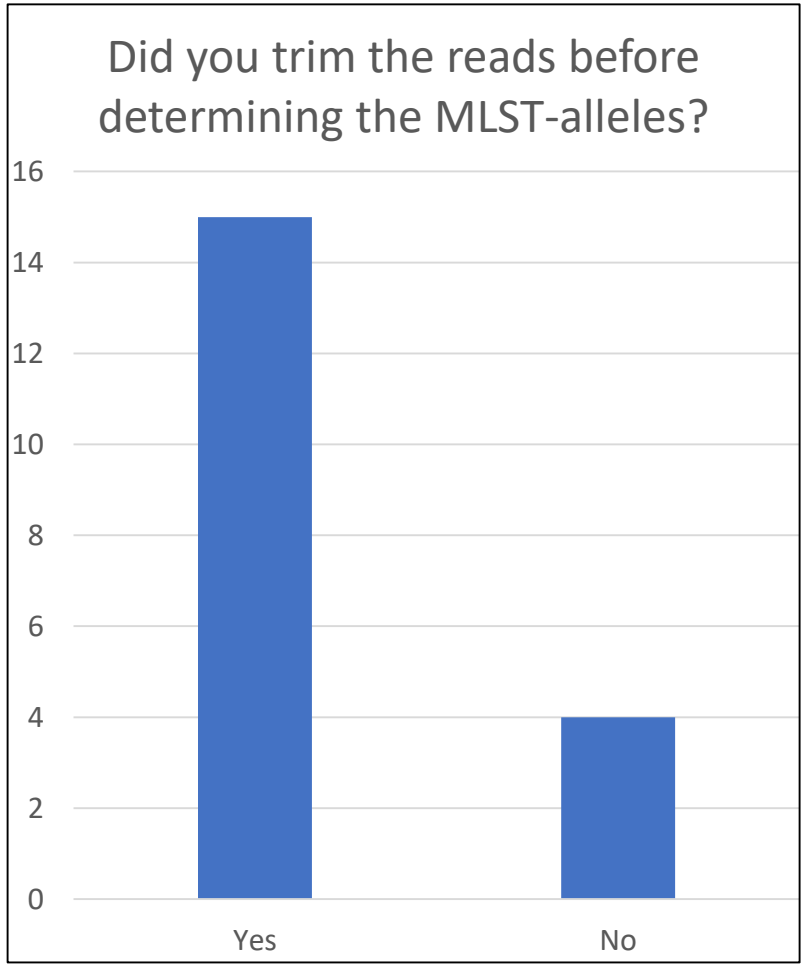




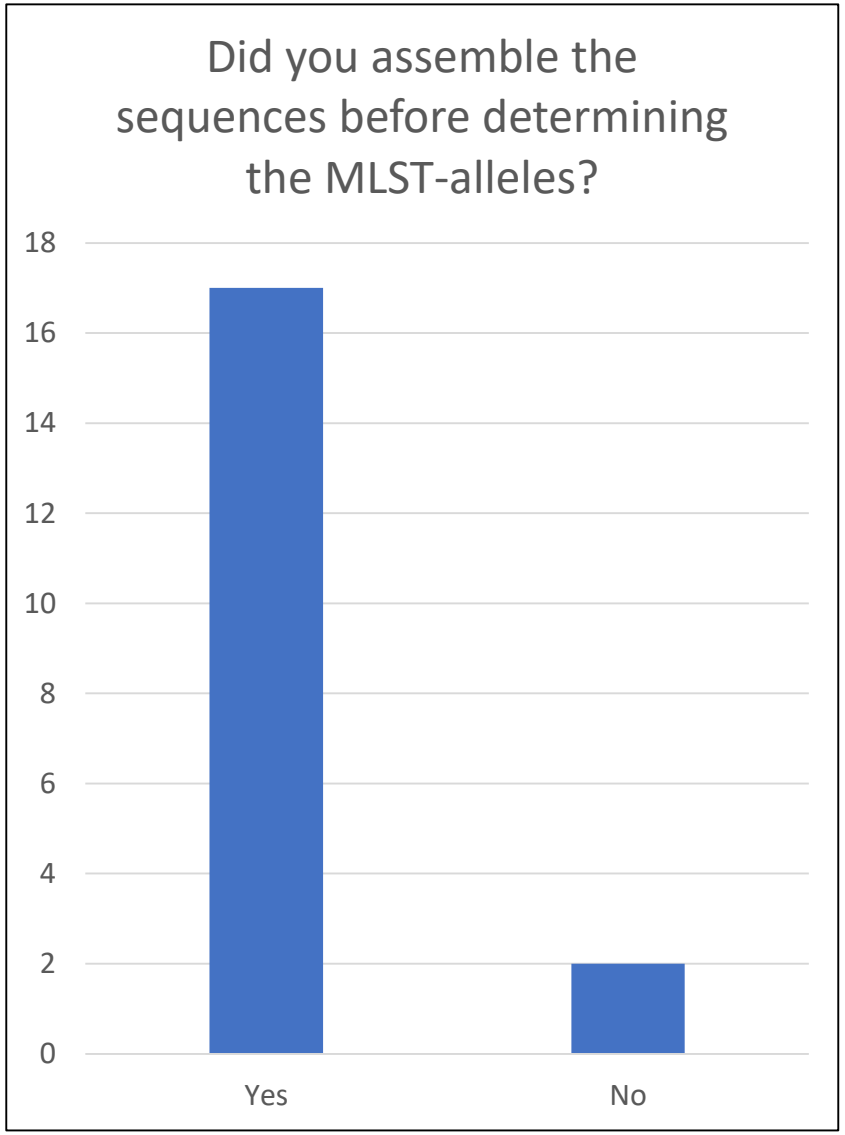
Procedures used by participants, WGS



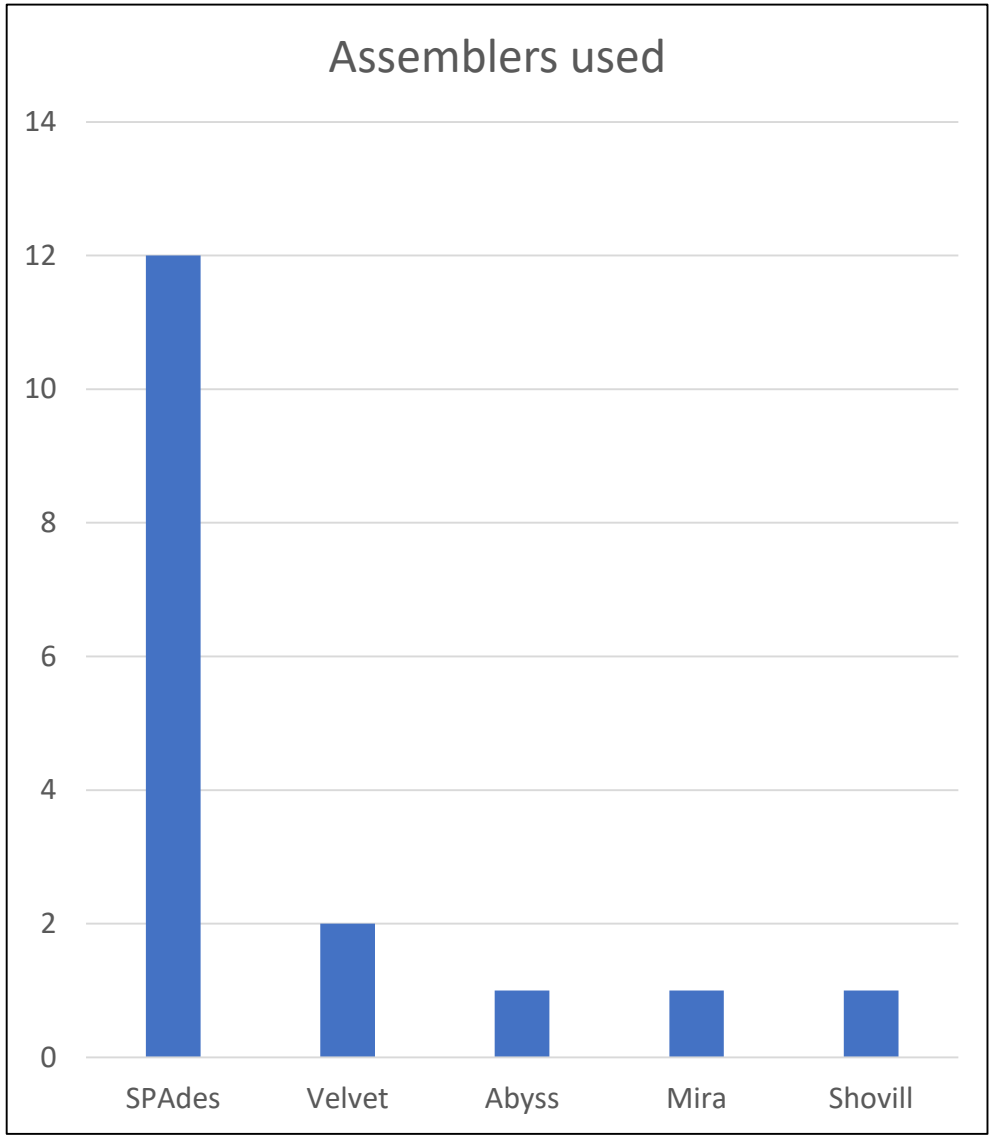
Procedures used by participants, WGS

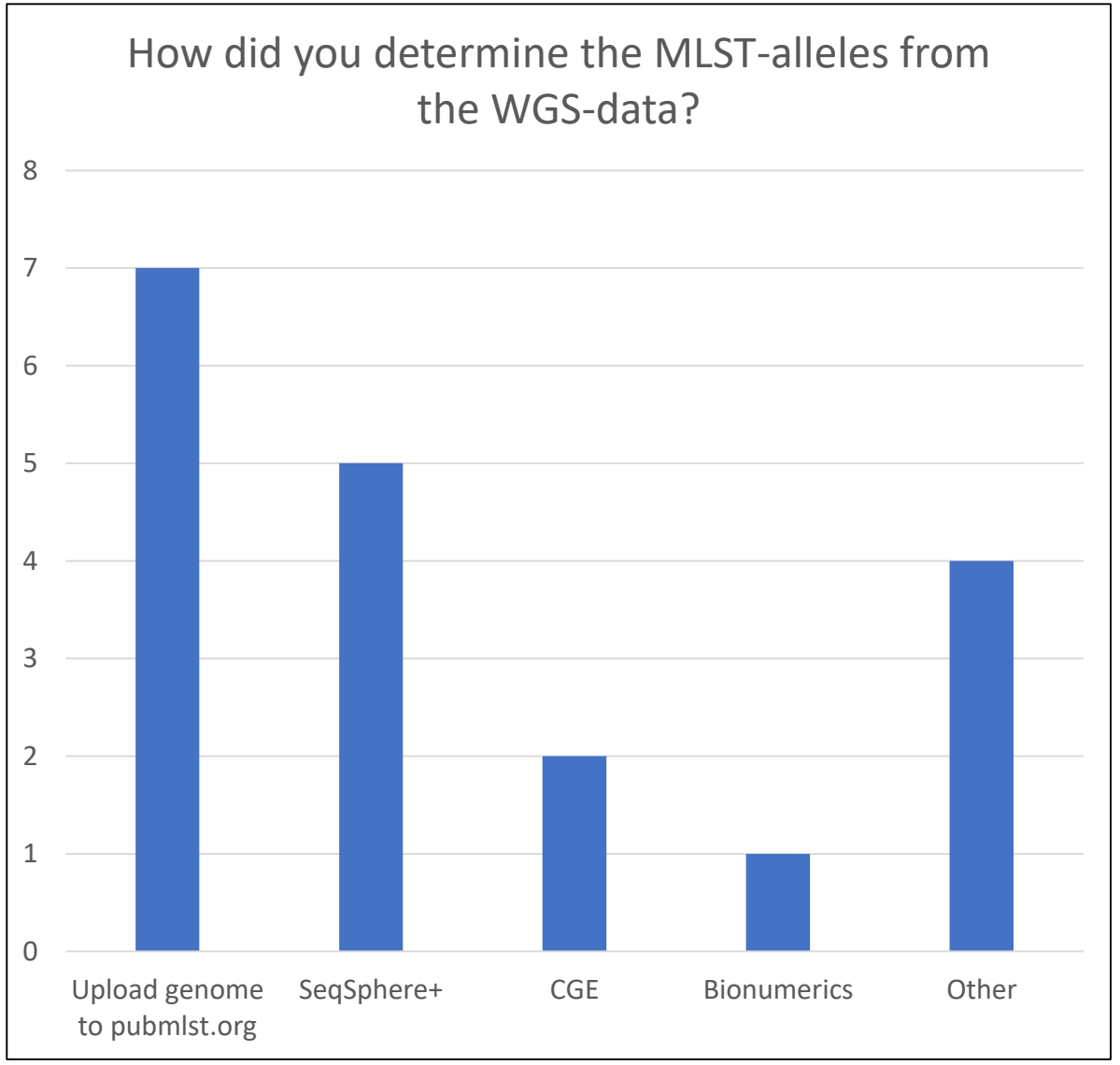


Majority used Trimmomatic



Procedures used by participants, WGS





RESULTS, WGS

No of labs	Correct ST (out of 8)	
17	8	
2	6	mix-up of samples (?) in both cases

NB! The IonTorrent S5 lab had to complement their data with Sanger on the *tkt* gene due to homopolymeric regions

We analysed the assemblies sent in by the 2 labs and the correct STs could be determined from them.
This means that all WGS labs sent in data that gave the correct ST (except IonTorrent)
Human error seems to be the cause - operators must have mixed up samples at reporting or analysis.

OPTIONAL CLUSTER ANALYSIS, WGS

cg/wgMLST

Instead of 7 genes – uses the entire core or whole genome.

Optimal for surveillance

SNP

Analyse individual single nucleotide polymorphisms.

The highest level of resolution.

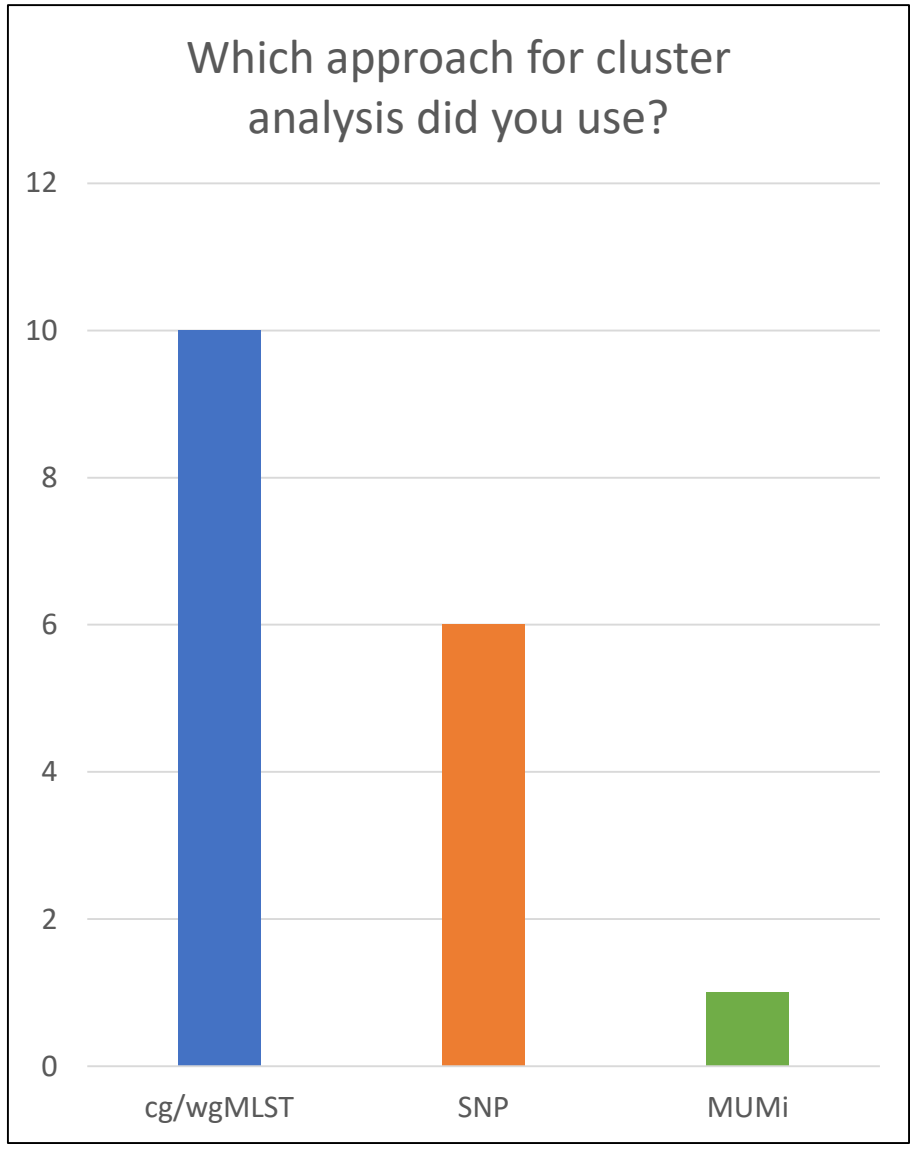
17/19 labs performed cluster analysis

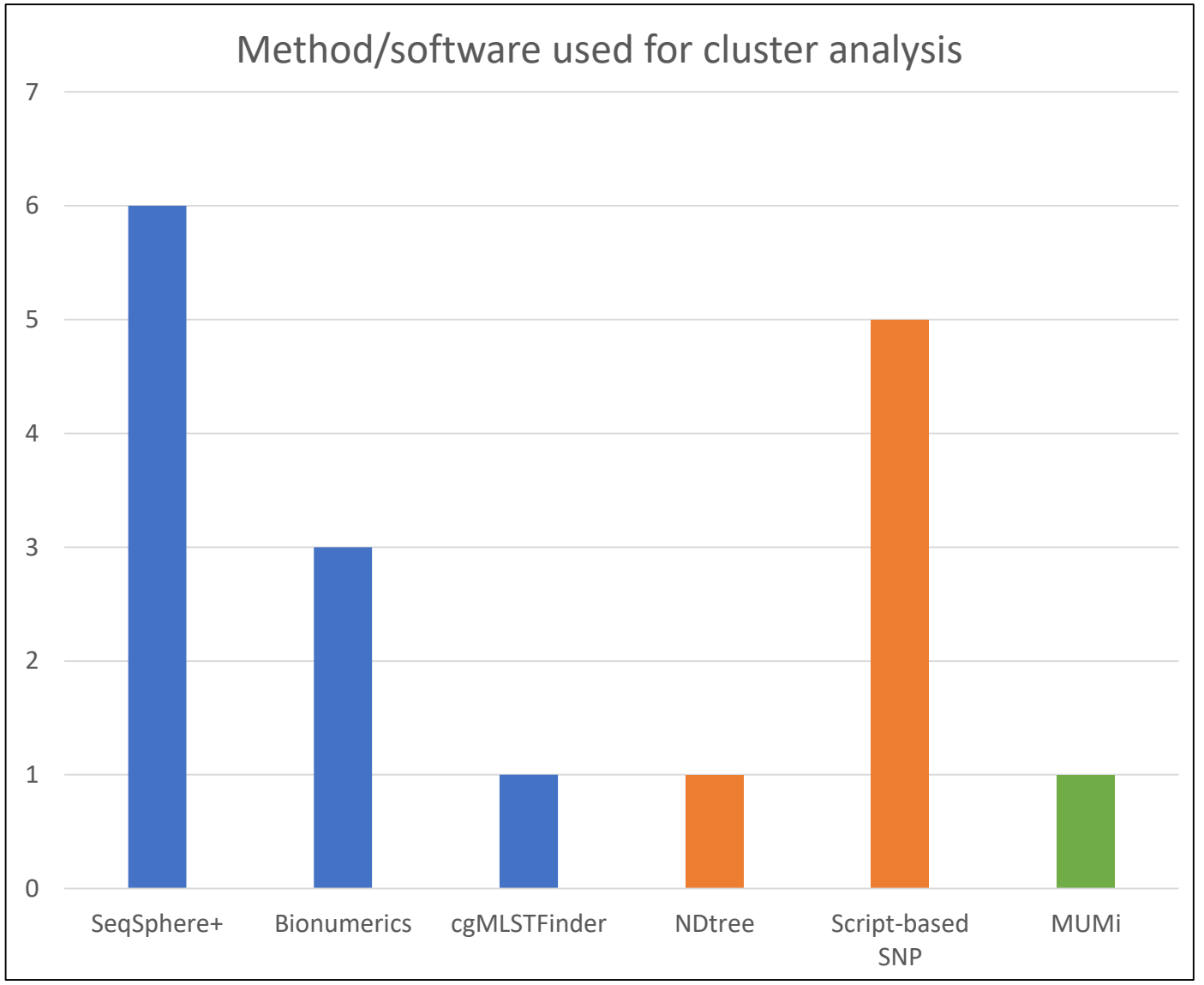
Participants were asked to upload raw sequence data, assembled data and images that supports the cluster analysis interpretations.

We have analysed the data using wgMLST to visualise and evaluate the results.

- Proposed method for comparing isolates across countries (Nadon et al. 2017)
- Majority of labs have used MLST approach
- SNP-comparisons might be performed for the report/paper

Procedures used by participants, WGS





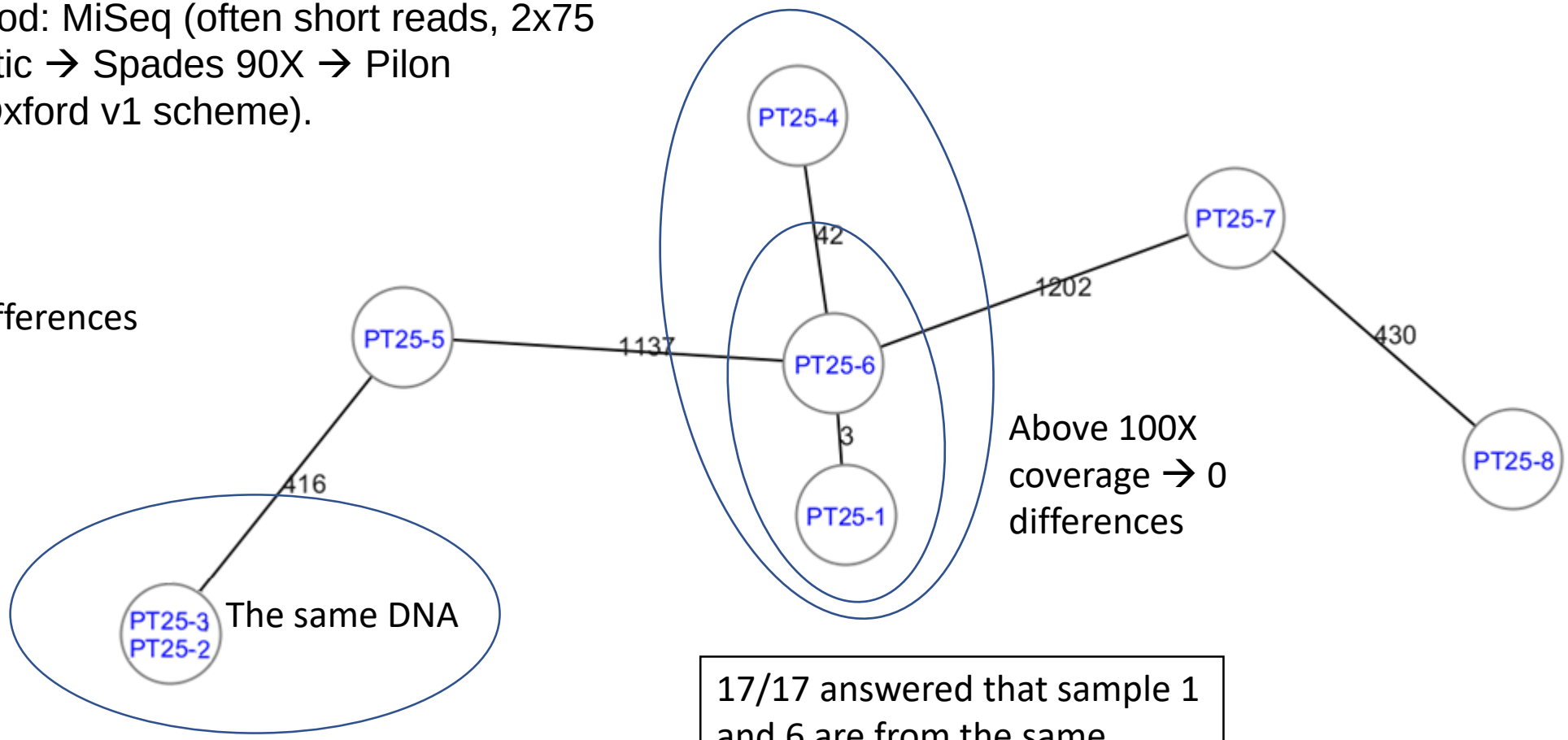
Selection of strains for PT25

- Isolates chosen from previous outbreaks and surveillance – not reference strains
- Samples had to be very diverse to make ST-determination interesting which means that the clusters do not contain many samples.
- Intention was to have two clusters - One simple (same DNA) and one with a more difficult interpretation.

Cluster analysis results and discussion

- Our primary method: MiSeq (often short reads, 2x75 bp) → Trimmomatic → Spades 90X → Pilon →SeqSphere+ (Oxford v1 scheme).

Numbers = Allele differences
PT25-X = Sample



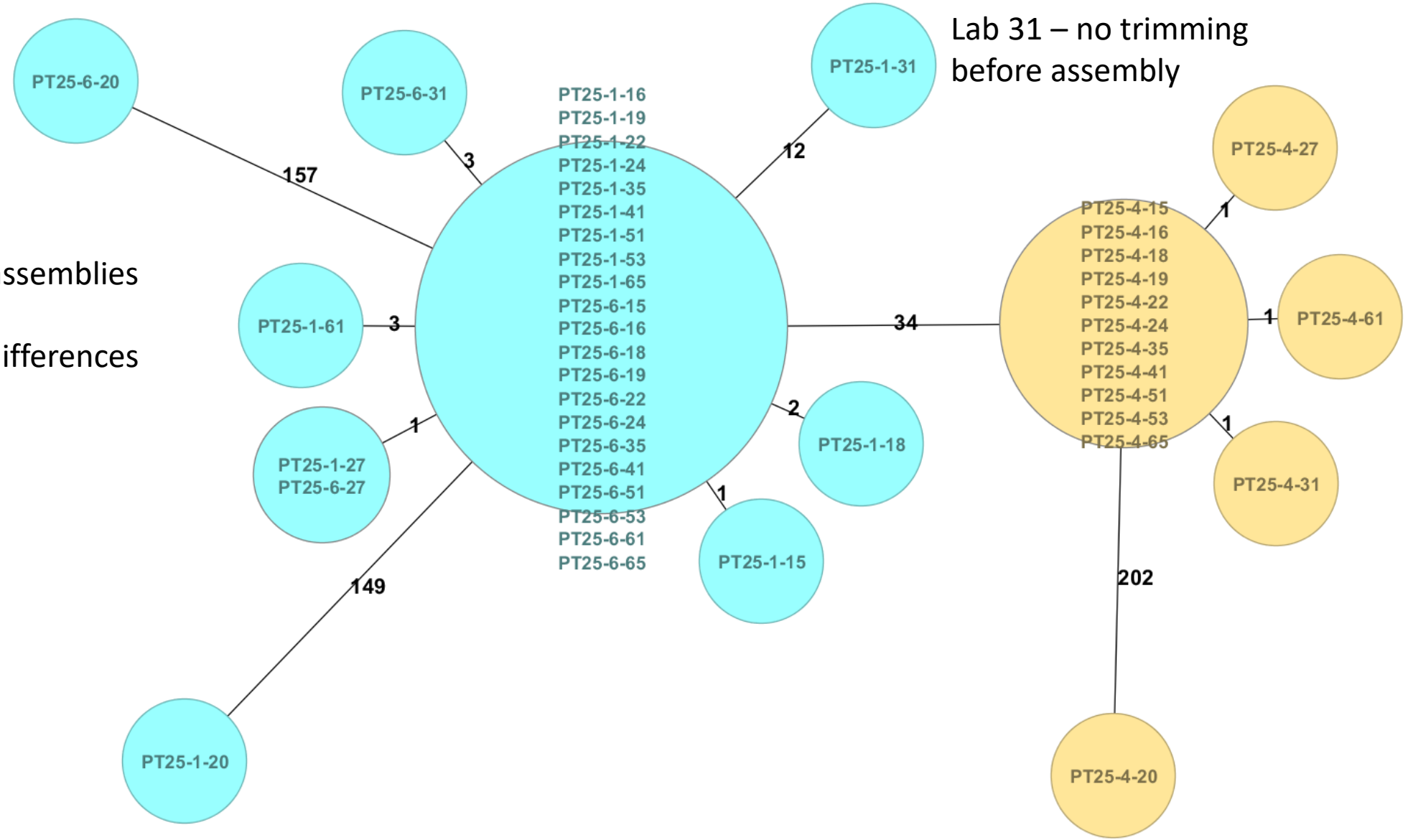
17/17 labs answered that sample 2 and 3 are from the same source

17/17 answered that sample 1 and 6 are from the same source.
4/17 answered that sample 4 is also from that source

Cluster analysis results and discussion

Cluster 1,4,6 NRL-assemblies

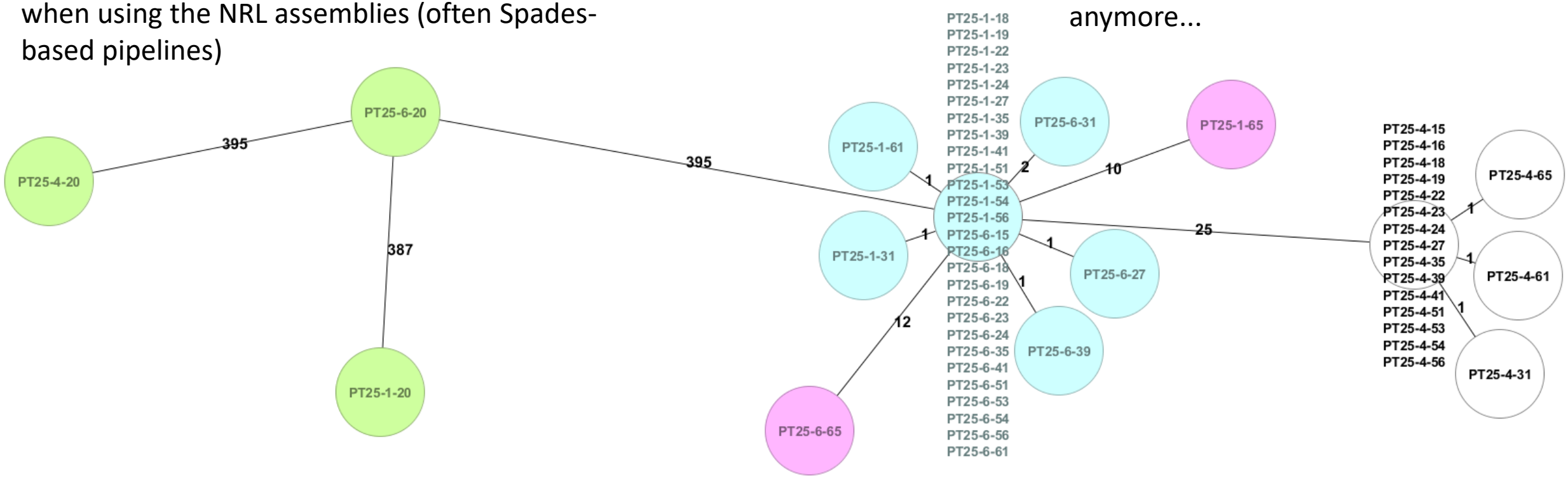
Numbers = Allele differences
PT25-X-Y = Sample



Cluster analysis results and discussion

Differences between our Spades-pipeline and when using the NRL assemblies (often Spades-based pipelines)

However, Lab 31 does not anymore...

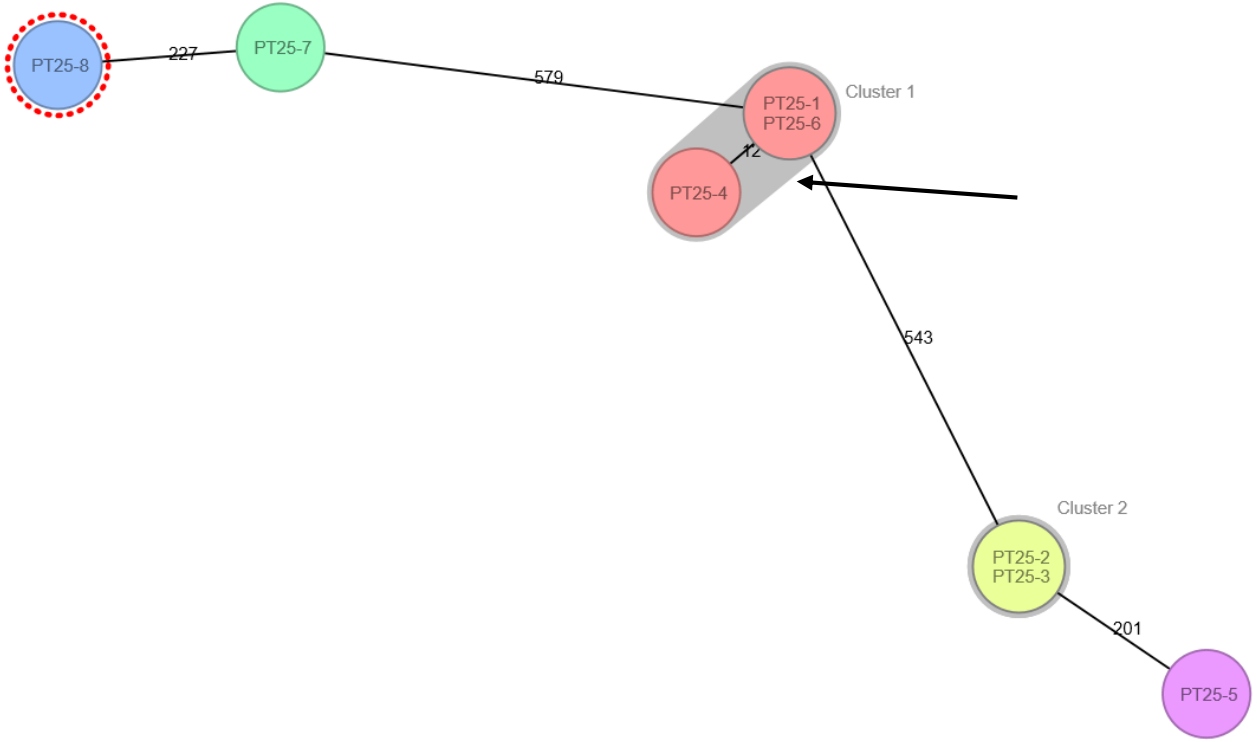


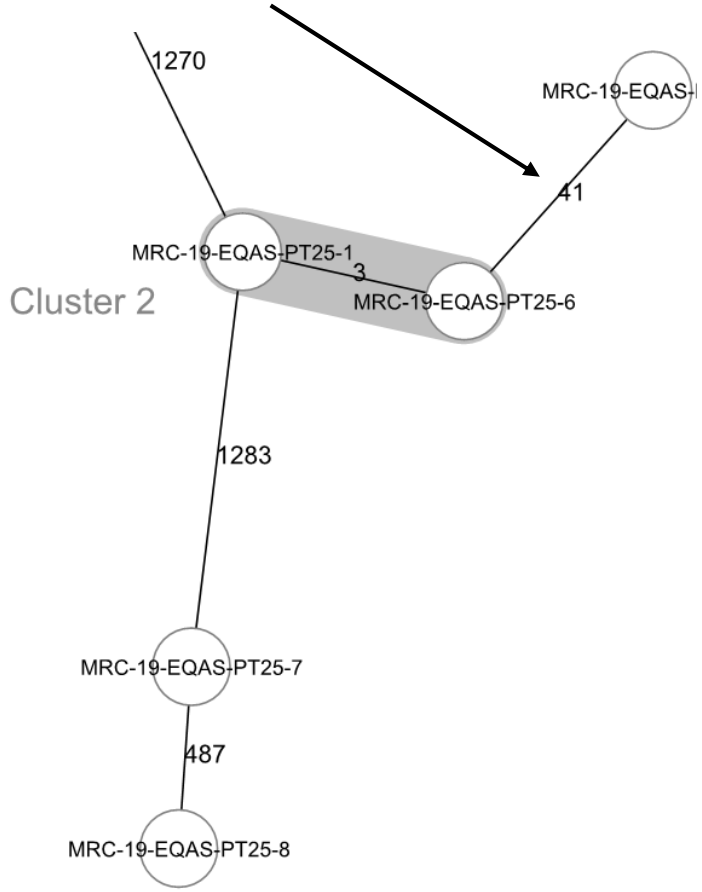
Twice the number of allele differences – IonTorrent needs special settings/software to assemble well.

Lab 65 shows many false alleles. Low quality data

Cluster analysis results and discussion

Difference in using only core compared to core + accessory

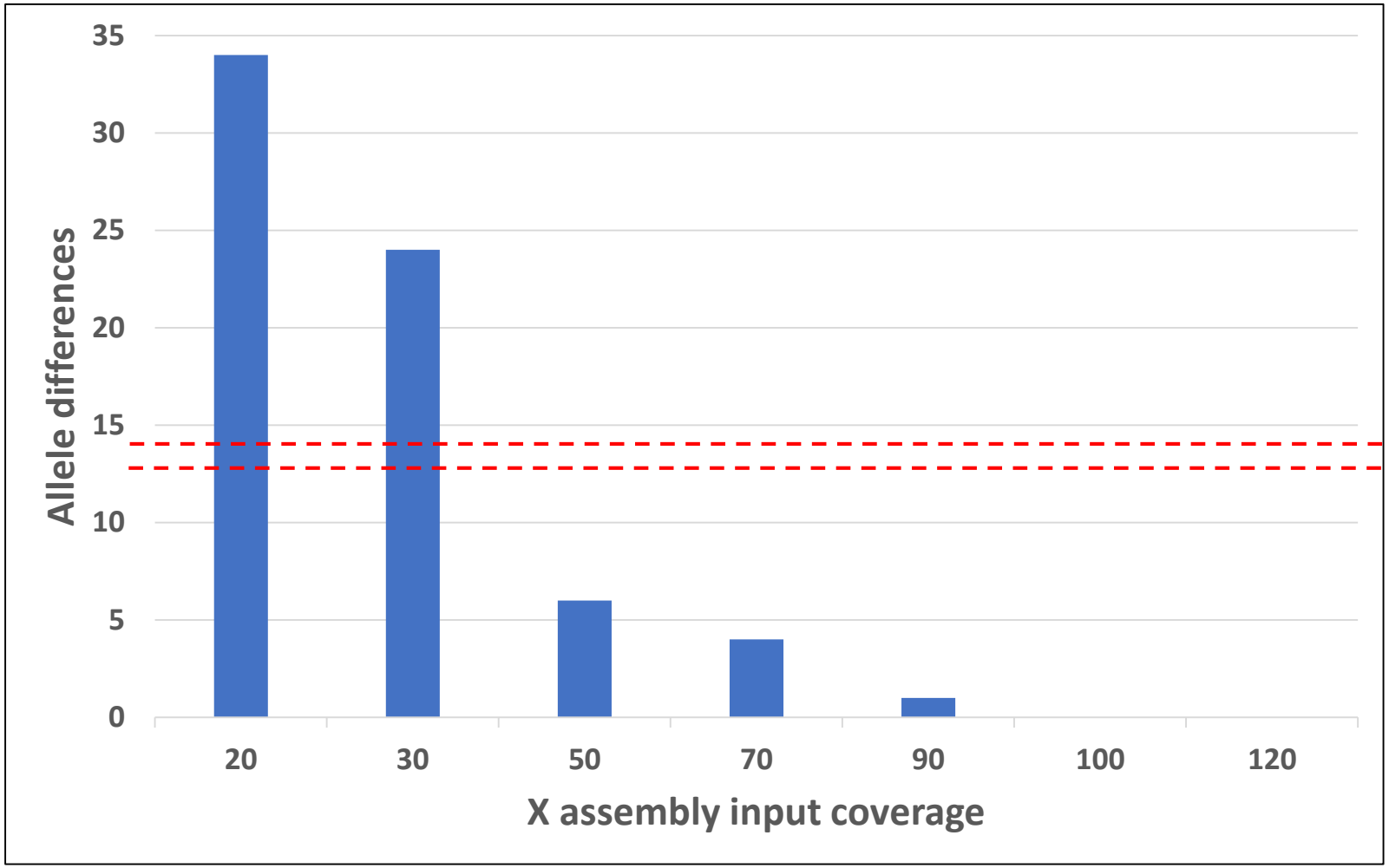




Task Templates: C. jejuni/coli cgMLST v1.3, C. jejuni/coli MLST v1.1
C. jejuni/coli cgMLST Complex Type / Cluster-Alert distance: 13
Comparison Table Retrieval: Campylobacter [unstored]
Comparison Table created: 24-May-2019 09:53 (v5.0.0_(2018-04))
Ridom SeqSphere+ MST for 8 Samples based on 644 columns, pairwise ignoring missing values
Distance based on columns from C. jejuni/coli cgMLST:cgMLST (637), C. jejuni/coli MLST (7)
For citing correctly in publications the tools used for this analysis see menu Help | Citations.
Cluster distance threshold: 13

Task Templates: C. jejuni/coli cgMLST v1.3, C. jejuni/coli MLST v1.1, C. jejuni/coli Accessory v1.2, C. jejuni/coli_flaA
C. jejuni/coli cgMLST Complex Type / Cluster-Alert distance: 13
Comparison Table Retrieval: Campylobacter [unstored]
Projects: Campylobacter (Campylobacter jejuni/coli)
Comparison Table created: 15.04.2019 15:10 (v5.1.0_(2018-06))
Ridom SeqSphere+ MST for 8 Samples based on 1595 columns, pairwise ignoring missing values
Distance based on columns from C. jejuni/coli cgMLST:cgMLST (637), C. jejuni/coli Accessory (958)
For citing correctly in publications the tools used for this analysis see menu Help | Citations.
Cluster distance threshold: 13

What happens in cgMLST with lower coverage?



”Cluster alert” is 13 in SeqSphere+

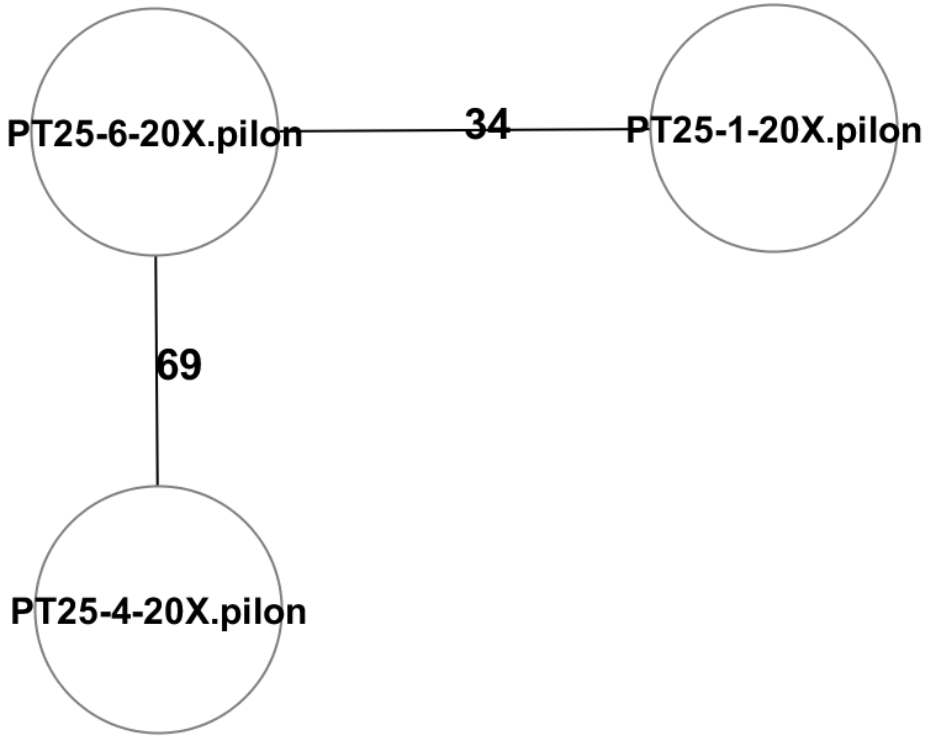
”Relatedness threshold” is 14 for *C. jejuni* (A.C.Schürch et al. 2018)

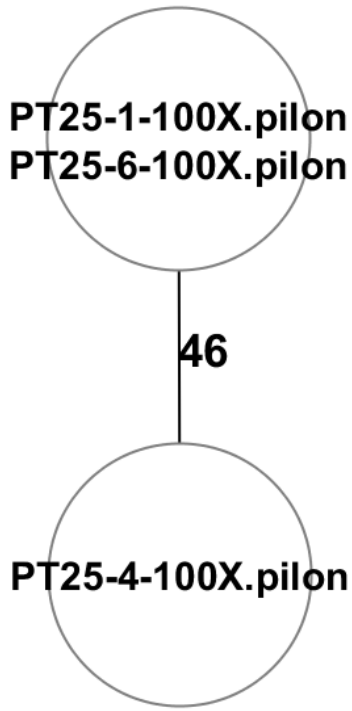
And, 444 alleles (33% of targets) fewer between 20X and 120X

However nr 1: Strain dependant

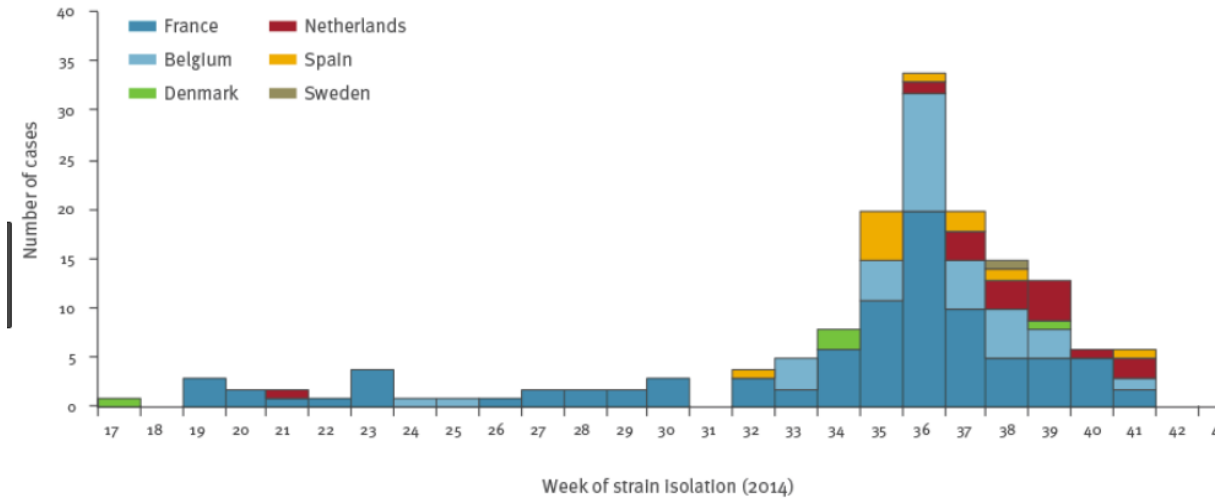
Samples 1 and 6

However nr 2: this adverse effect can be minimised using only high-quality reads. MiniSeq 34X – 50X, 2 x 150 bp > no such effect

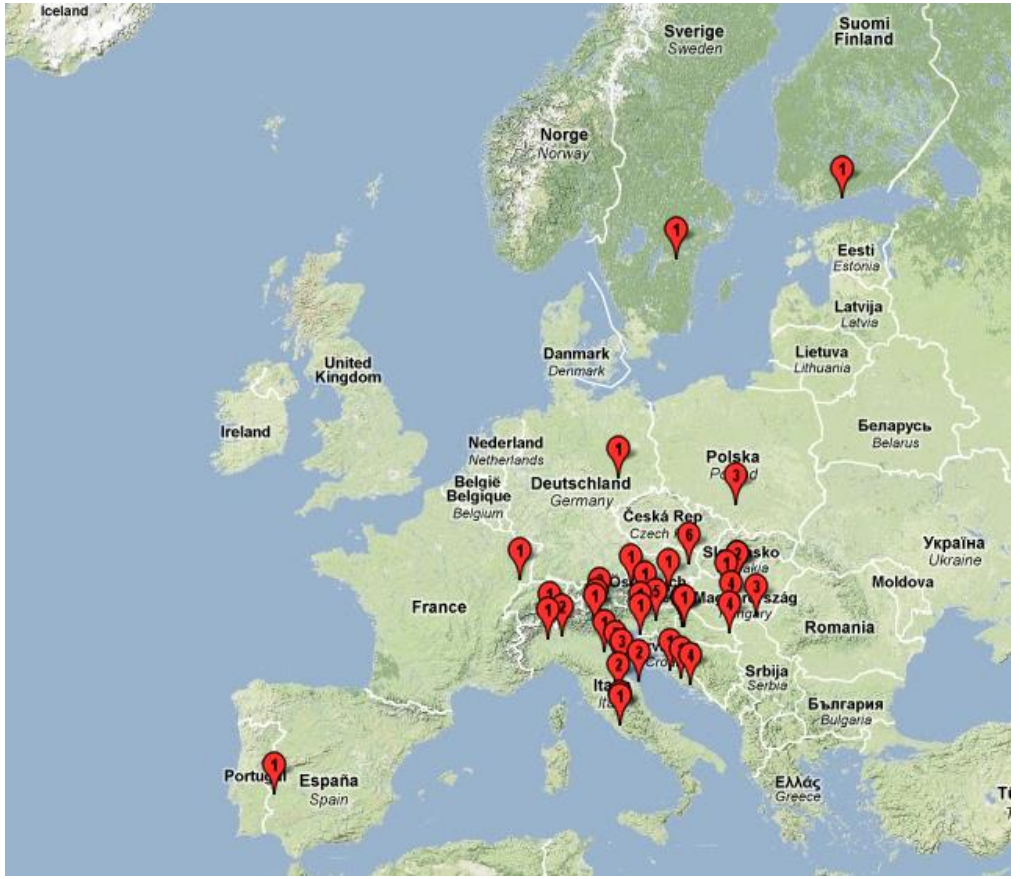

PT25-6-20X.pilon
34
PT25-1-20X.pilon
69
PT25-4-20X.pilon


PT25-1-100X.pilon
PT25-6-100X.pilon
46
PT25-4-100X.pilon

Would it work?



A multi-country outbreak (19 countries)



Summary

- For Sanger, there is room for improvement
- For WGS, if not counting operator mix-ups when reporting, the results were very good
- If the analysis had been performed at one location using one pipeline – the WGS raw data would have produced 100% accurate STs and clustering results for 18/19 labs (IonTorrent data impossible or needs experience the EURL lacks?).
- When setting up a WGS pipeline you should keep in mind that coverage together with read quality can have a big impact for MLST-based methods.