Inter EURLS Working Group on NGS:

Proficiency Tests on Next Generation Sequencing

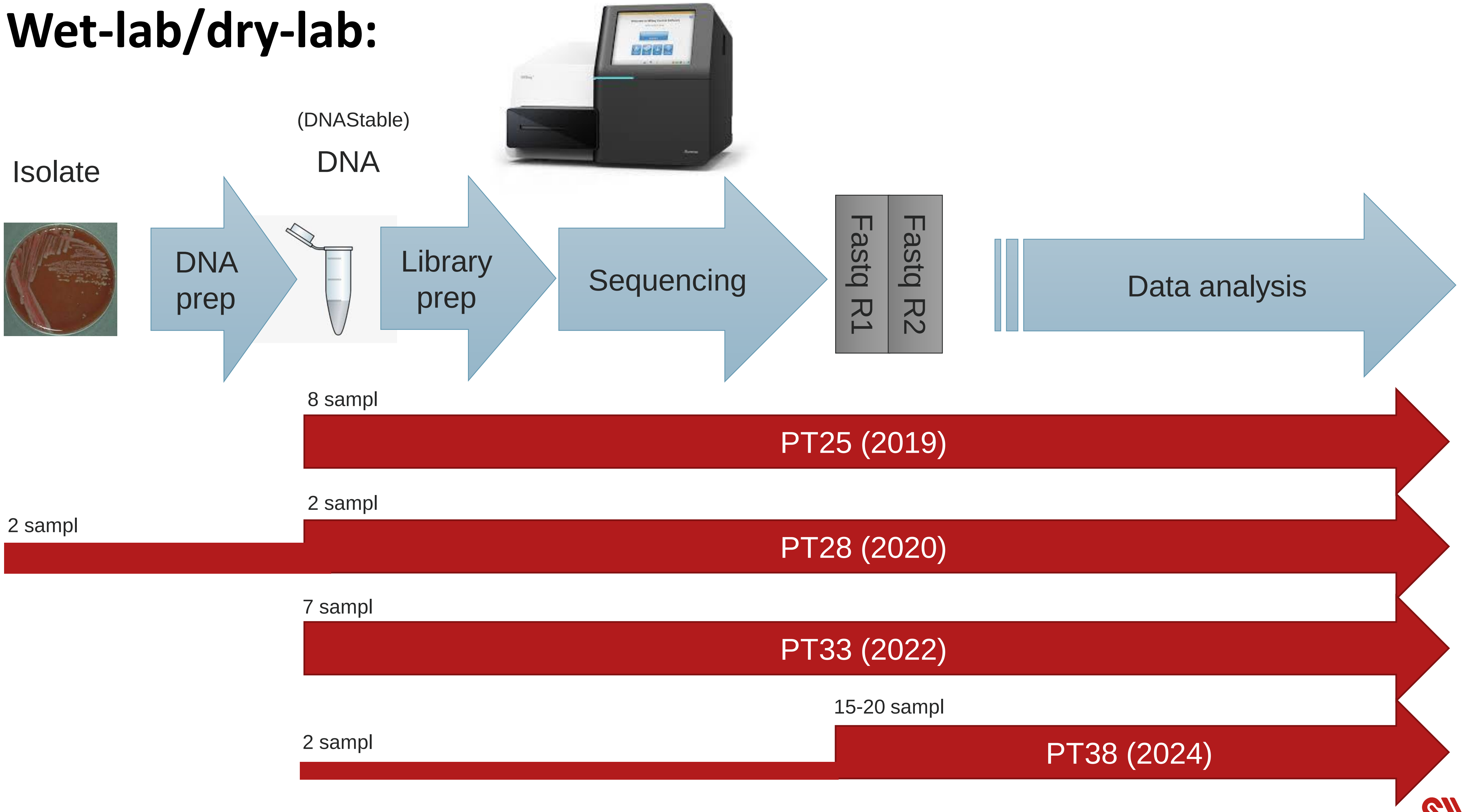
Whole Genome Sequencing and Cluster analysis of *Campylobacter*

Bo Segerman
EURL-*Campylobacter*
bo.segerman@sva.se

Sep 2023

Co-funded by the European Union

Wet-lab/dry-lab:



Isolate

DNA prep

(DNAStable)

DNA

Library prep

Sequencing

Fastq R1

Fastq R2

Data analysis

8 sampl

PT25 (2019)

2 sampl

2 sampl

PT28 (2020)

7 sampl

PT33 (2022)

15-20 sampl

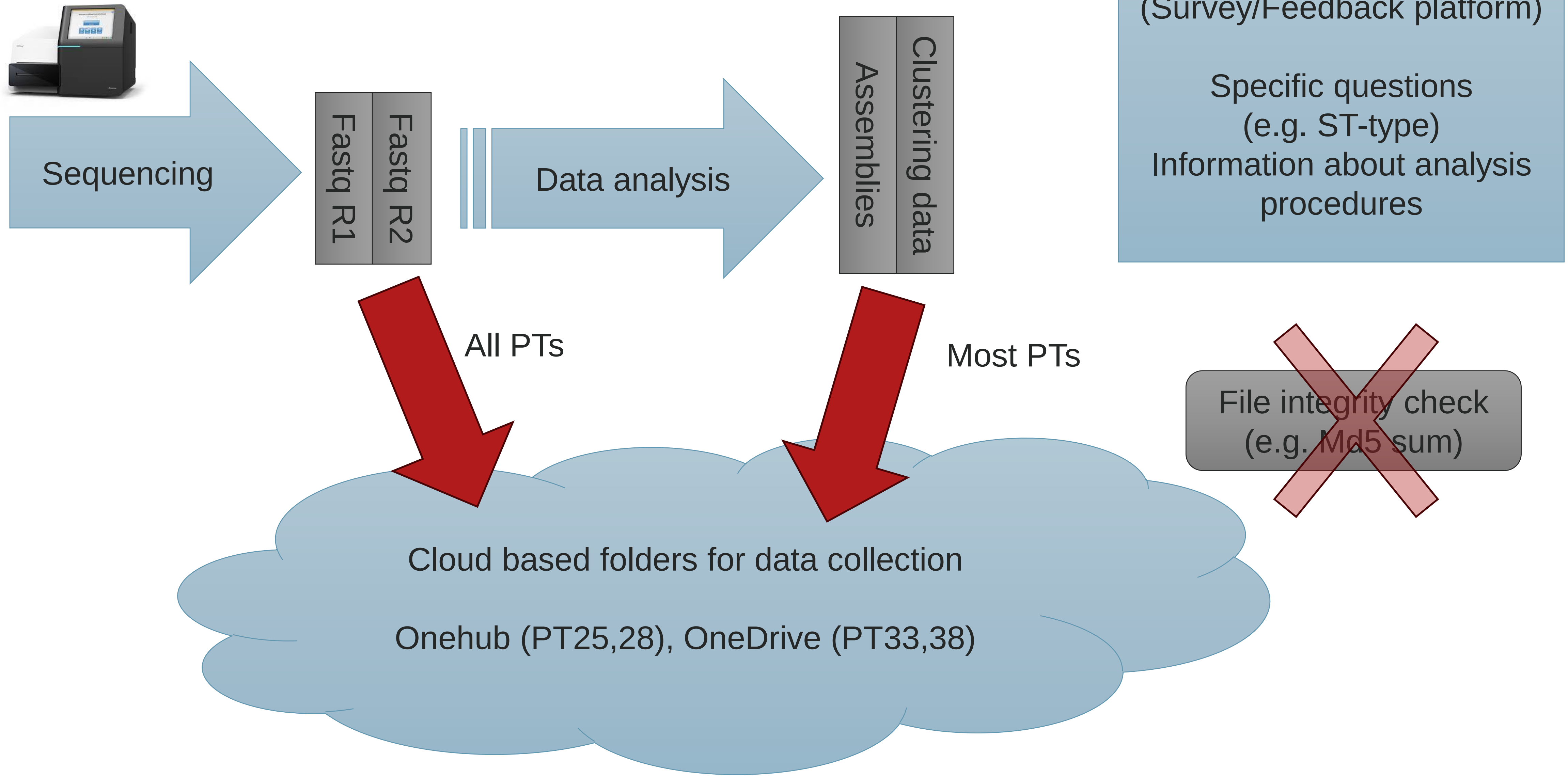
2 sampl

PT38 (2024)

Collection of results:

Questback:
(Survey/Feedback platform)

Specific questions
(e.g. ST-type)
Information about analysis procedures



Sequencing

Fastq R1

Fastq R2

Data analysis

Assemblies

Clustering data

All PTs

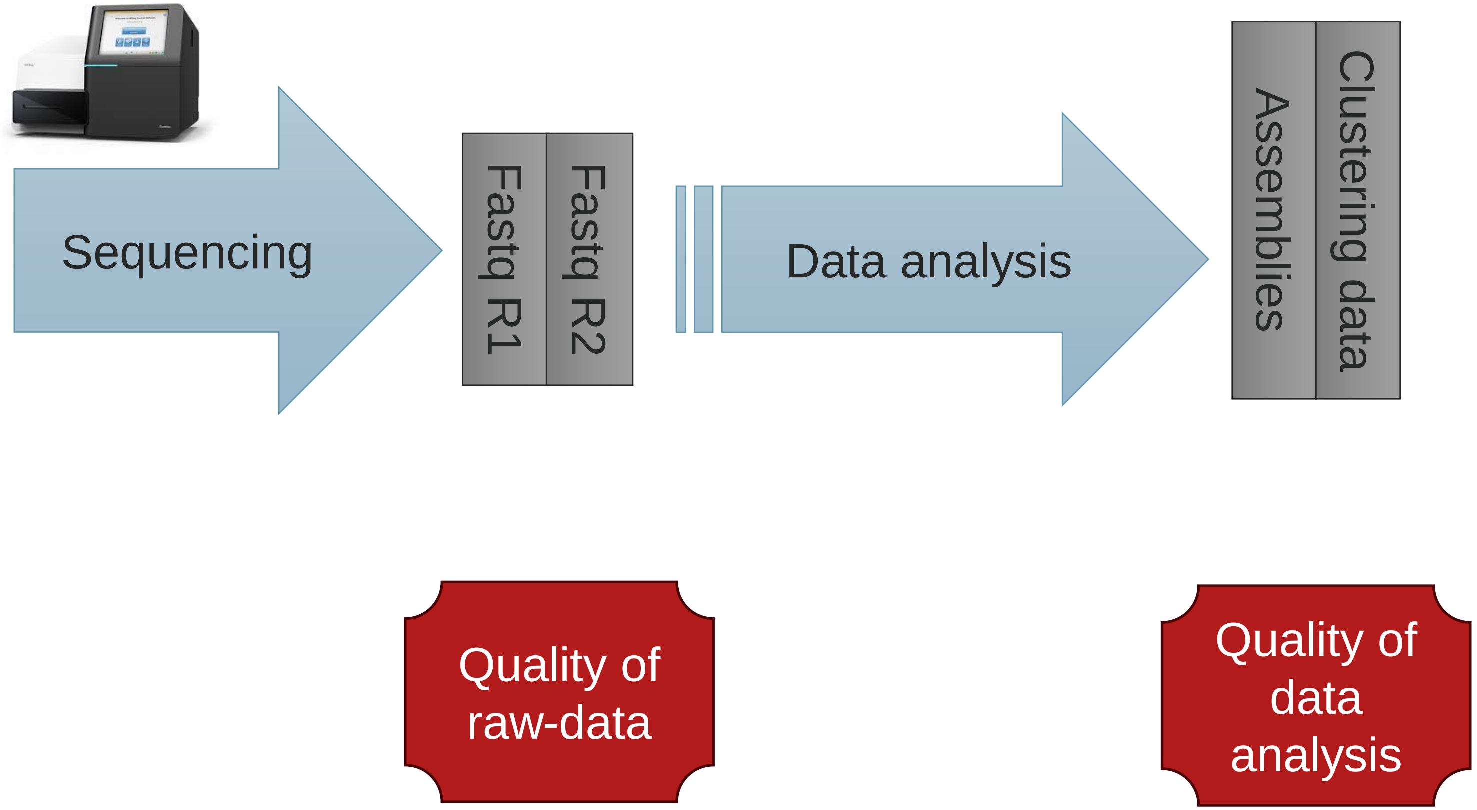
Most PTs

Cloud based folders for data collection

Onehub (PT25,28), OneDrive (PT33,38)

~~File integrity check~~
~~(e.g. Md5 sum)~~

Results analysis and report:



Sequencing

Fastq R1

Fastq R2

Data analysis

Assemblies

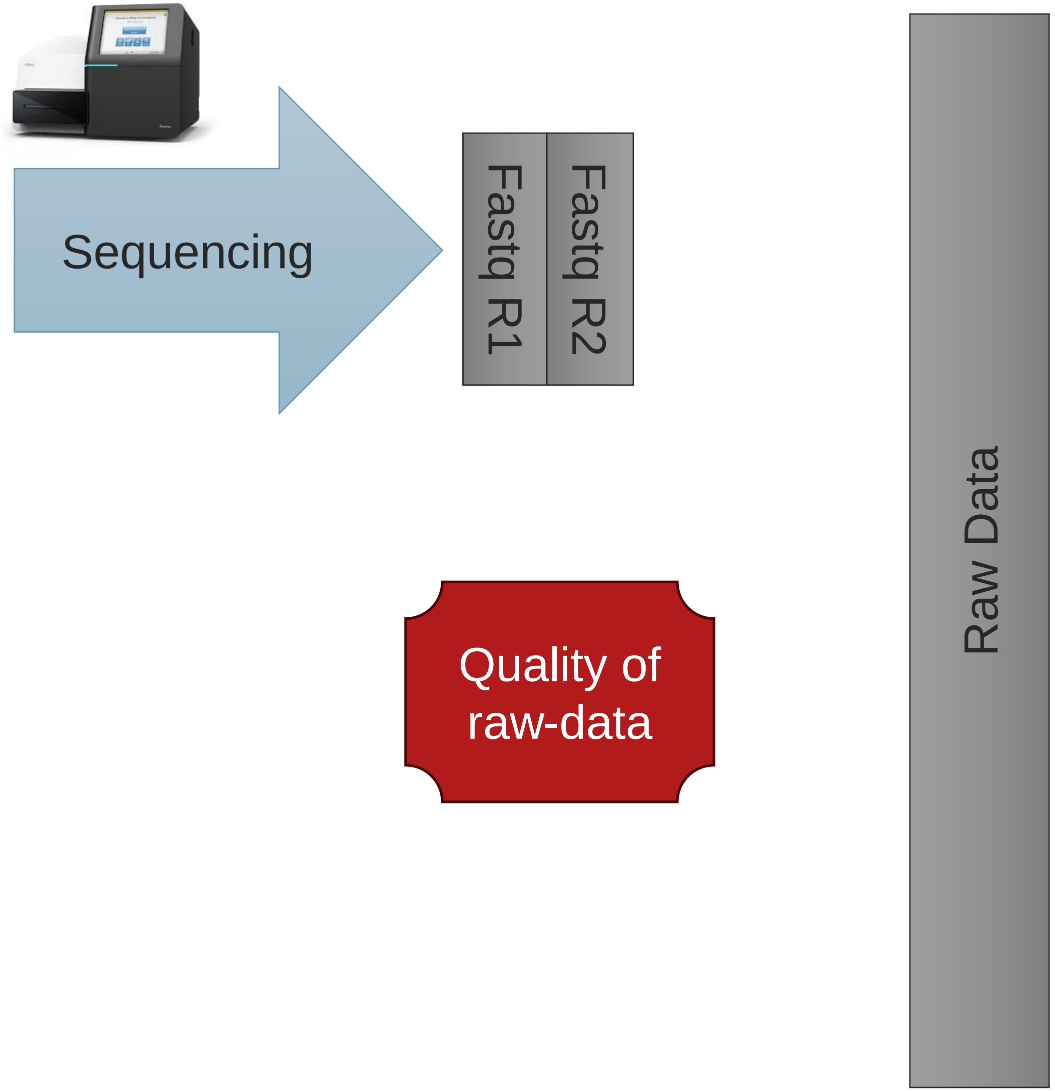
Clustering data

Quality of raw-data

Quality of data analysis

High quality Reference genome
(long read sequencing (ONT) + Illumina)

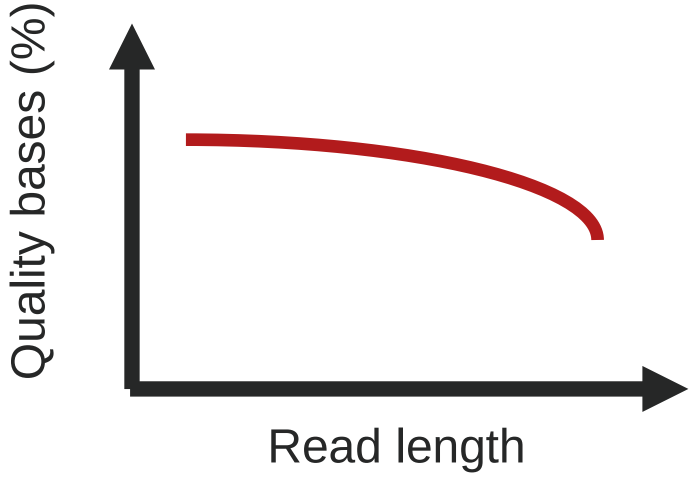
Results analysis and report:



Properties of raw data that affects quality meassurement

- Adapter content (fragment length)
- Prior quality trimming <= Affects results
- Read length (sequencing cycles)

Quantify high quality bases Q30 baser (%)



Different threshold for different read lengths

Results analysis and report:

Sequencing

Fastq R1

Fastq R2

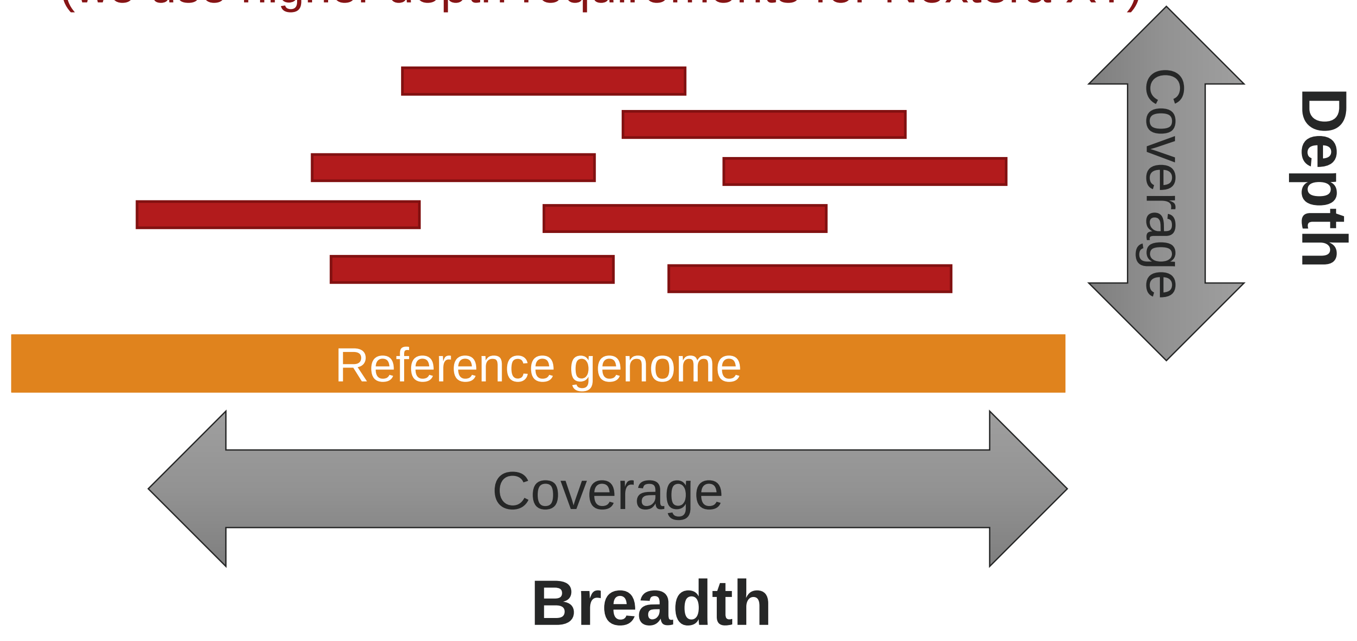
Quality of raw-data

Raw Data

- Coverage depth
- Coverage breadth
- Coverage fluctuations

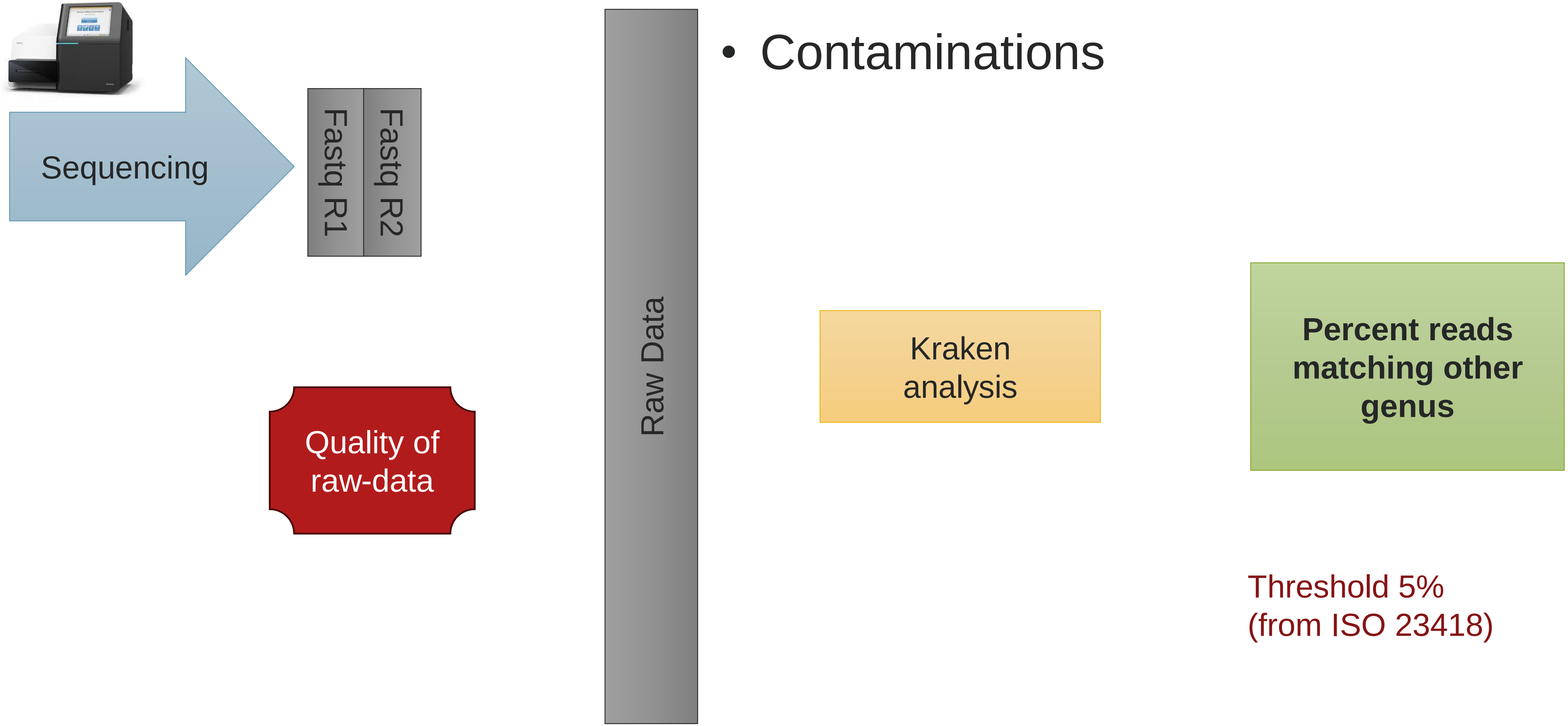
Quantity of data after trimming
(X times the reference genome size)

Nextera XT users have this problem
(we use higher depth requirements for Nextera XT)



Percent of the reference genome covered
(evaluated at a specific depth)

Results analysis and report:



- Contaminations

Sequencing

Fastq R1

Fastq R2

Raw Data

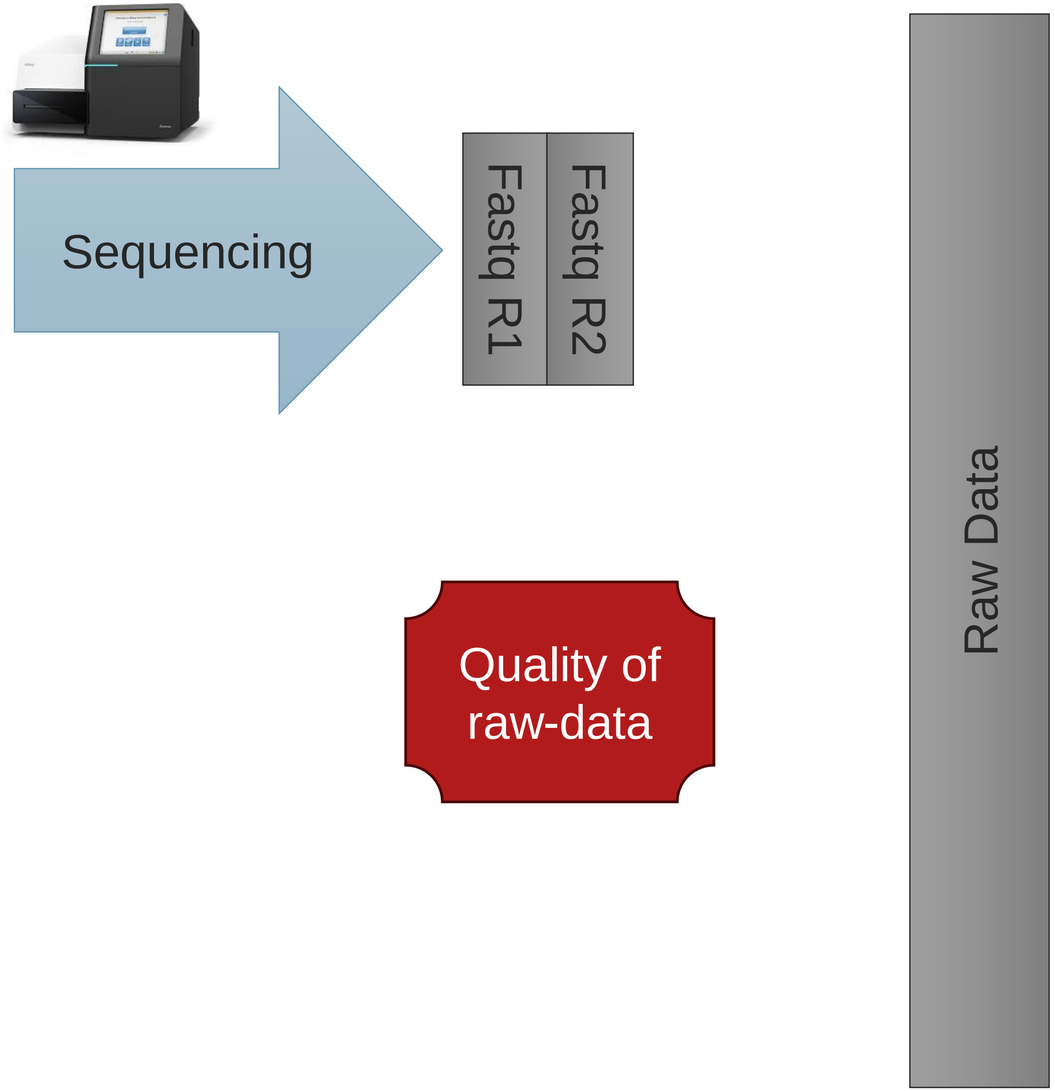
Quality of raw-data

Kraken analysis

Percent reads matching other genus

Threshold 5%
(from ISO 23418)

Results analysis and report:



Sequencing

Fastq R1

Fastq R2

Quality of raw-data

Raw Data

- GC deviation

Difference between average GC in reads and average GC in reference genome

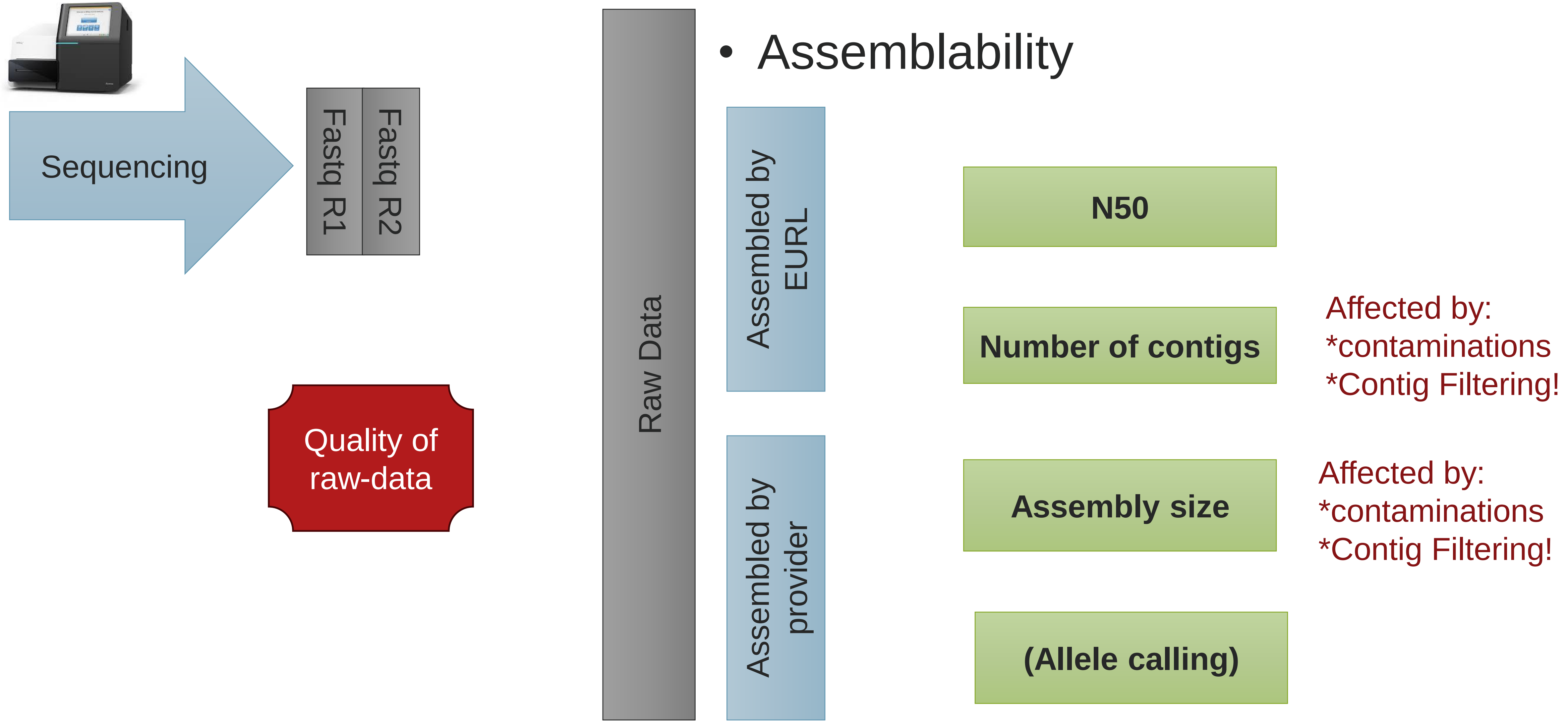
Threshold is 4%
(from ISO 23418)

Affected mainly by:

*Contamination (contaminant has other GC)

*GC bias in library prep kit (Nextera XT)

Results analysis and report:



Sequencing

Fastq R1

Fastq R2

Quality of raw-data

Raw Data

- Assemblability

Assembled by EURL

Assembled by provider

N50

Number of contigs

Affected by:
*contaminations
*Contig Filtering!

Assembly size

Affected by:
*contaminations
*Contig Filtering!

(Allele calling)

Results analysis and report:

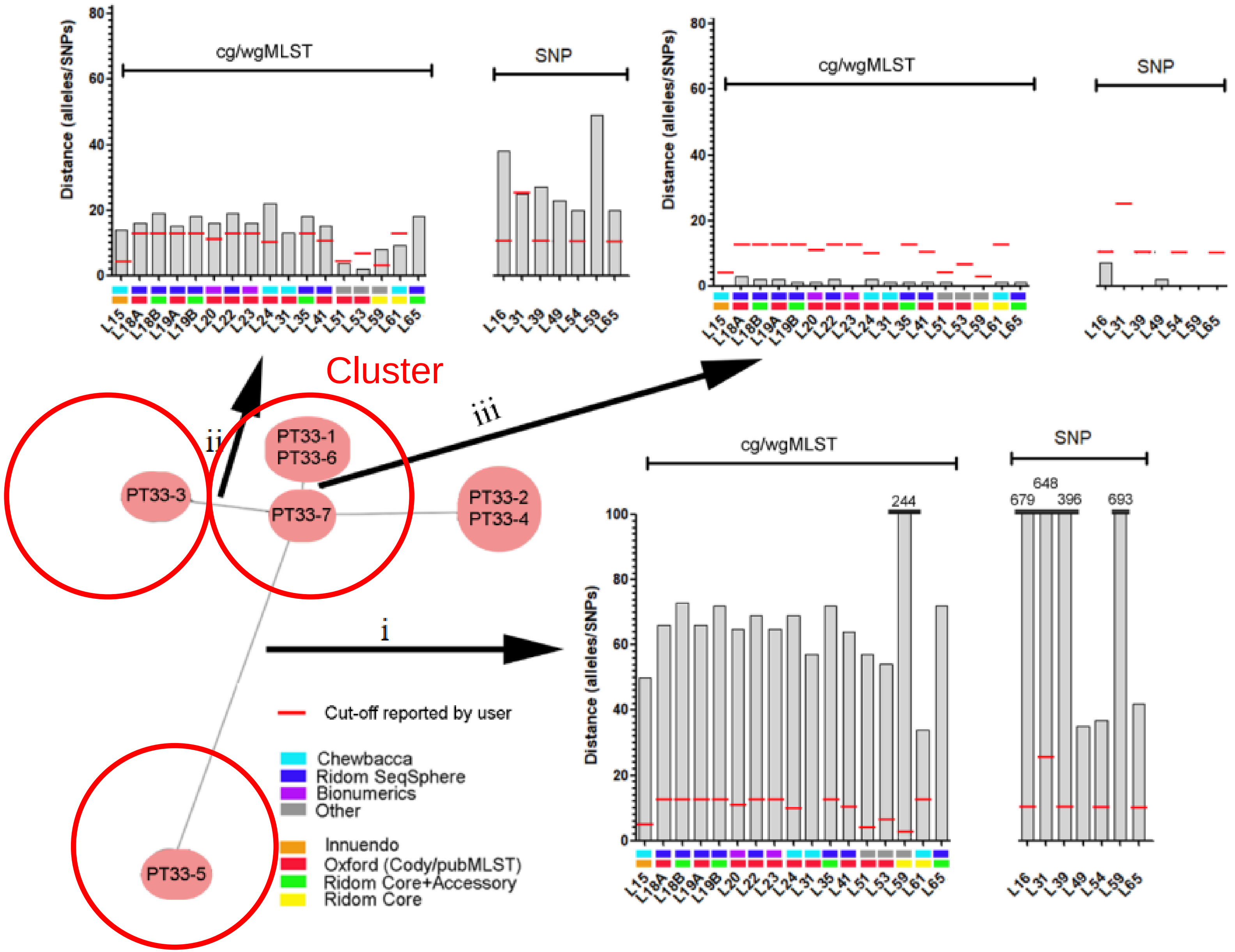
Data Heterogeneity:
Some labs use SNP, Some cgMLST/wgMLST
Different schemas…Different SNP pipelines

Data analysis

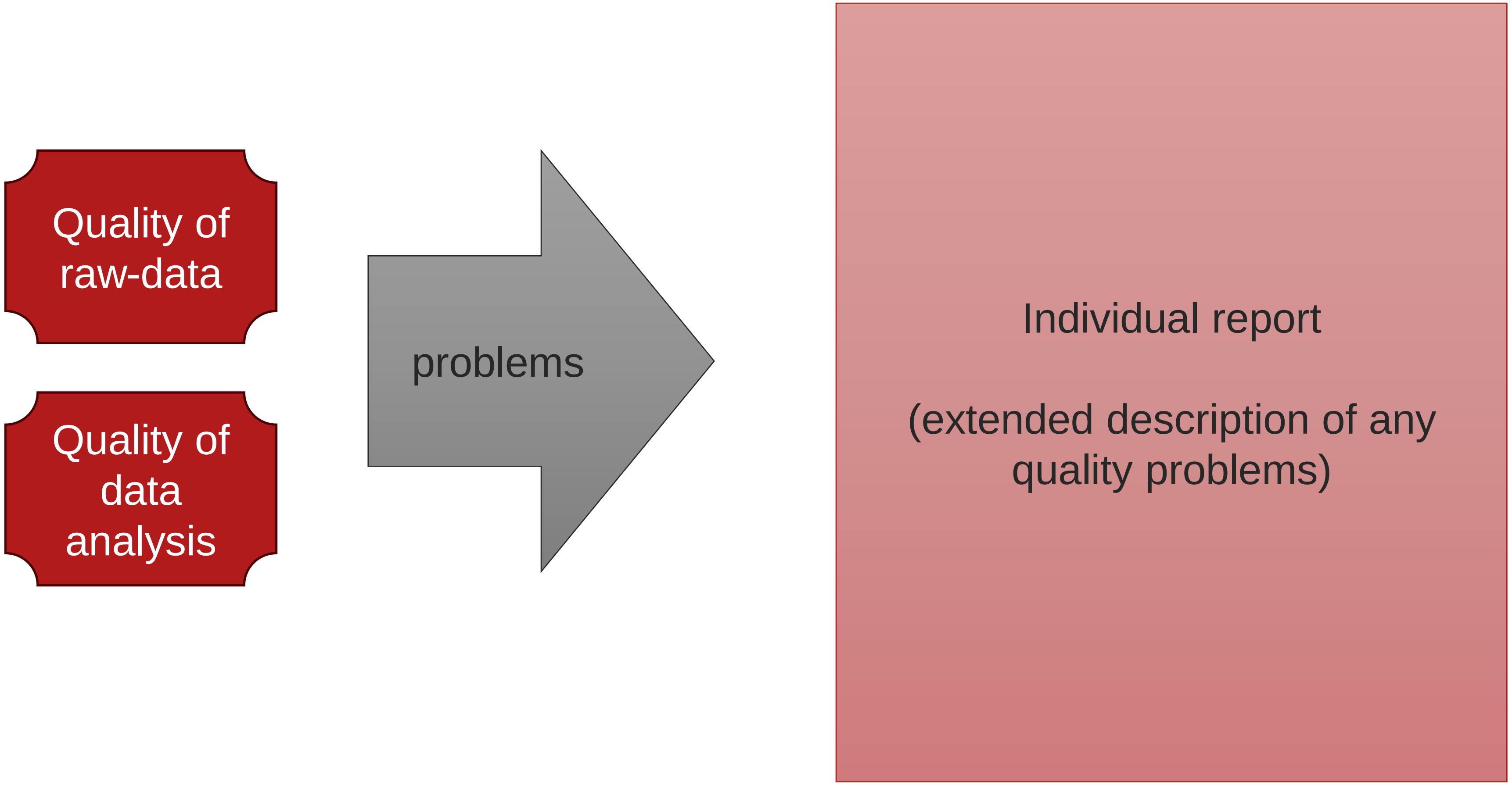
Quality of data analysis

Closer relationship

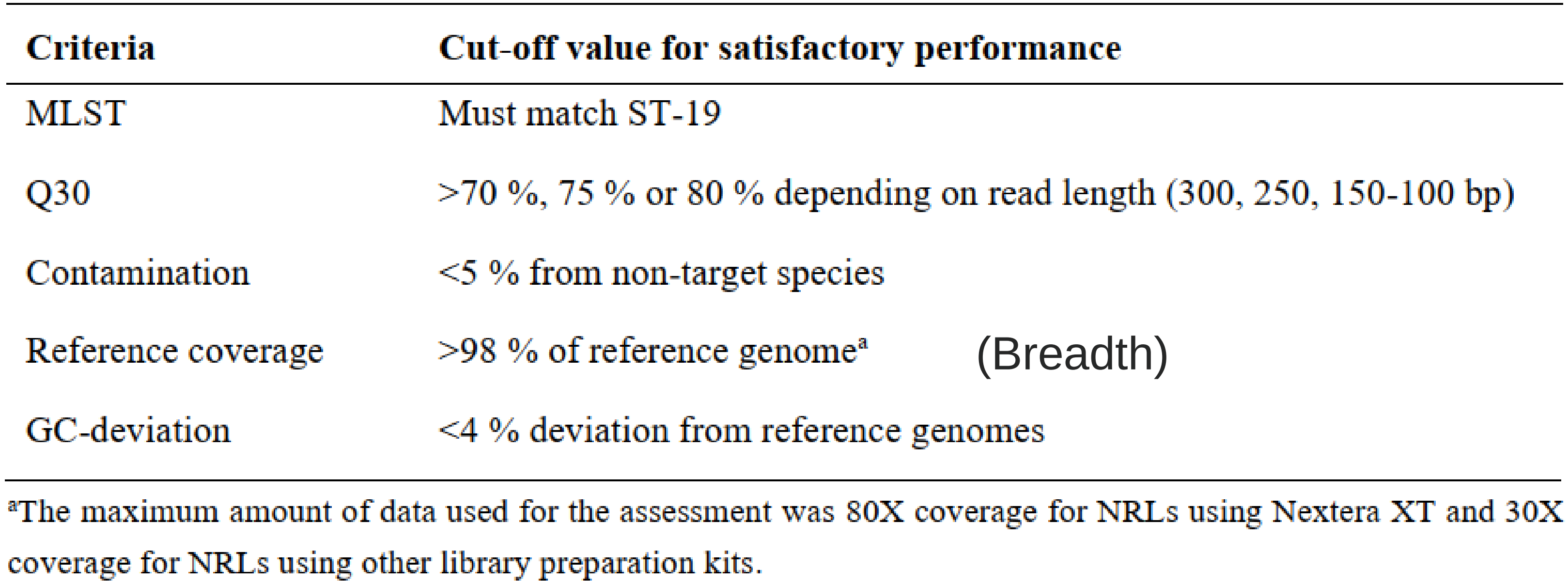
Distant relationship



Results analysis and report:



Performance assessment



Criteria	Cut-off value for satisfactory performance
MLST	Must match ST-19
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

[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.

(Breadth)

No overall scoring

Satisfactory / needs improvement

for each criteria

Clusters (Topology)

”X and Y are closest to Z”
”X is the most distant sample”

Lessons learnt

- Raw data QC parameters are affected by several factors
    - Read length (cycles)
    - Pre-made trimming and filtering steps
    - Library kit used (Nextera XT – coverage fluctuations)
- Different thresholds may need to be used depending on library prep kit / read length (cycles)
- Many QC measures are affected by several quality factors simultaneously
- Data analysis by EURL (comparability high, assesses quality of raw data, perhaps not optimized for the data)
- Data analysis by participant (comparability lower, assesses quality of data analysis)
- Clustering data is technically heterogenous and depend on context specific cutoff values