EURL-*Campylobacter*
Proficiency test 33

Whole genome sequencing and cluster analysis of *Campylobacter*

Ásgeir Ástvaldsson

EURL-*Campylobacter*

SVA, Sweden

27. April 2023

FWD AMR – RefLabCap Network meeting

Copenhagen, Denmark

Funded by
the European Union

EURL-*CAMPYLOBACTER*

Located at the National Veterinary Institute in Uppsala, Sweden

Network: 33 NRLs in member states, and 11 NRLs in third countries

Proficiency tests

- Annually for enumeration, detection and species identification of *Campylobacter* (ISO 10272-1 and ISO 10272-2).

- Every other year for **NGS**, next will be in 2024

Proficiency test 33

Whole genome sequencing and cluster analysis of *Campylobacter*

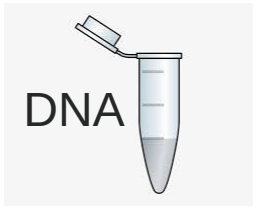
Objective

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

Purpose

- To help laboratories in the implementation of WGS and cluster analysis
- To test the joint capability of the network to solve a multi-country *Campylobacter* outbreak based on WGS data

EURL-*Campylobacter* PT 33



Lib. prep → Seq → Adapt rem → Q trim → ST type, Clustering analysis, AMR profile

7 samples

20 participants sent in:

- Raw data
- Assembly
- Cluster analysis images and distans matrices
- Questionnaire

PT33 divided into two parts

Sequence quality

Assessment of results
„Cut-off“ values defined for 5 criteria

- ST type: ”must match”
- Percent Q30: 70%,75%,80% (read length dependent)
- Contamination: <5%
- Reference coverage: 98% kmers
- GC content deviation: 4%

Cluster analysis

Assessment of results
Three statements used to capture topology

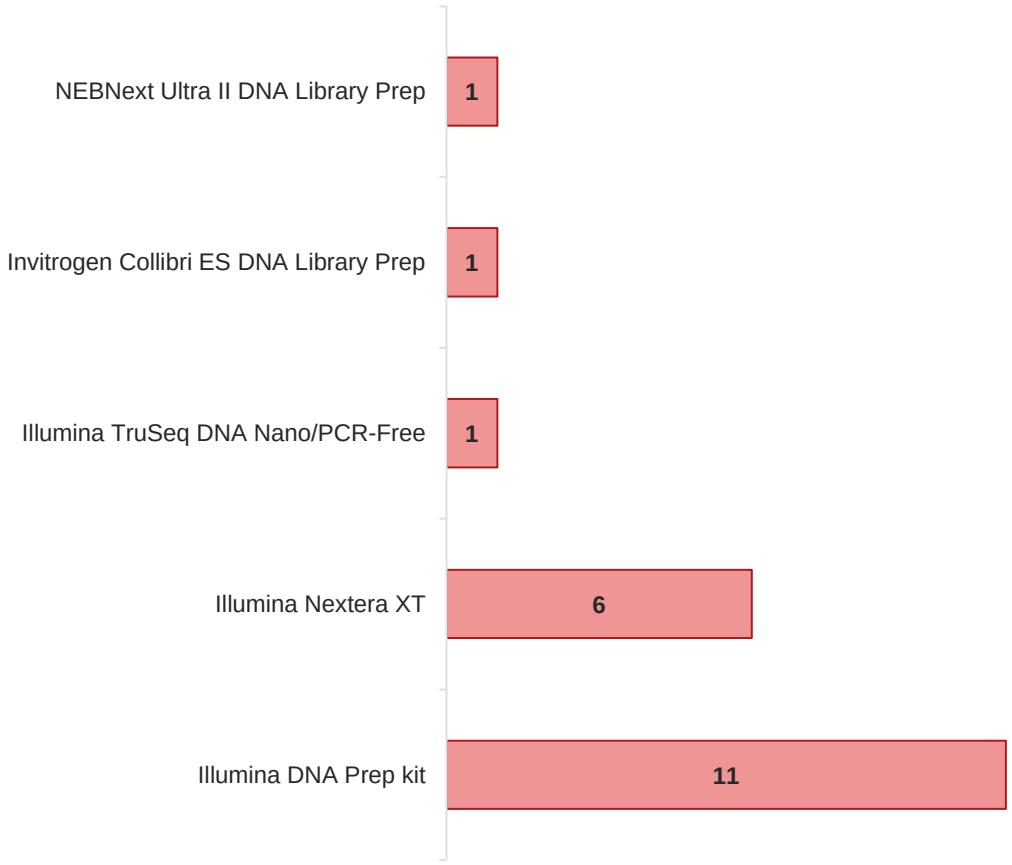
(i) “PT33-6 and PT33-7 are the two closest samples to PT33-1”
(ii) “PT33-4 is the closest sample to PT33-2”
(iii) “PT33-5 is most distant to the other samples

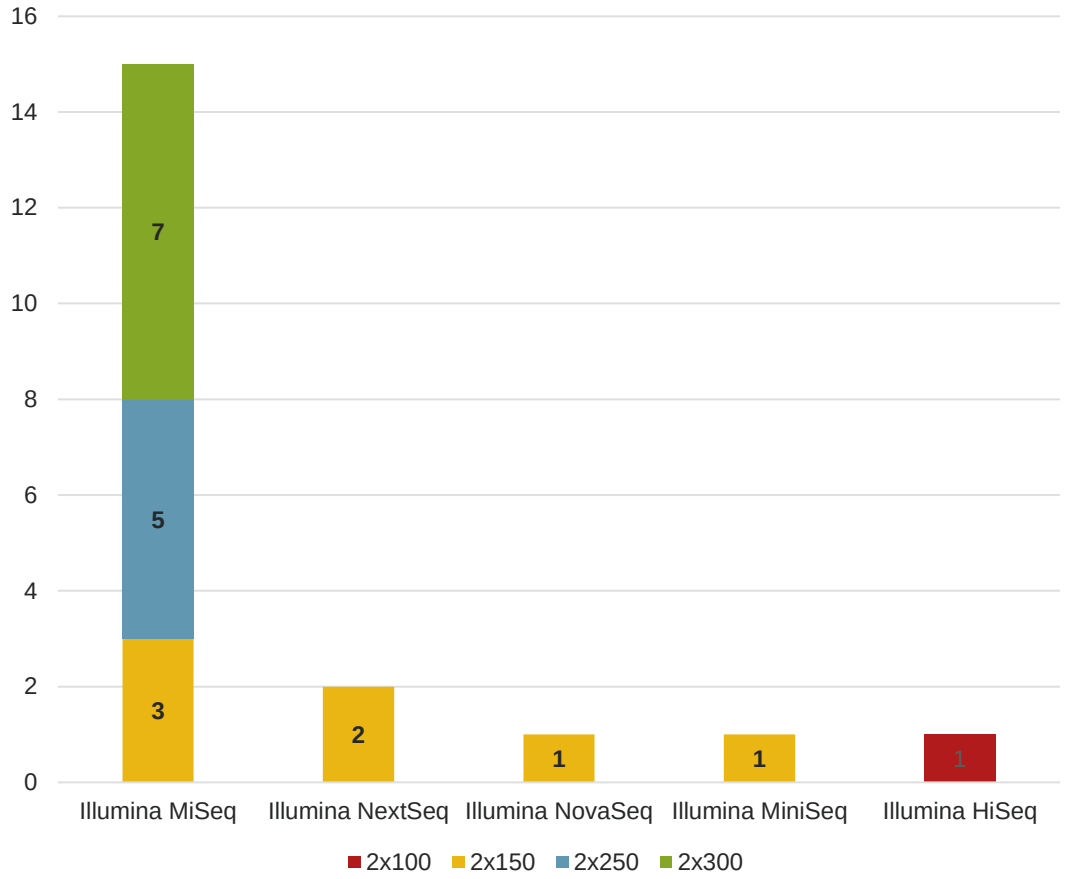
No overall performance criteria, but ”**satisfactory**” or ”needs **improvement**” for each criteria/statement

Each laboratory received individual report with results and comments on the data and possible improvements

Lessons learned

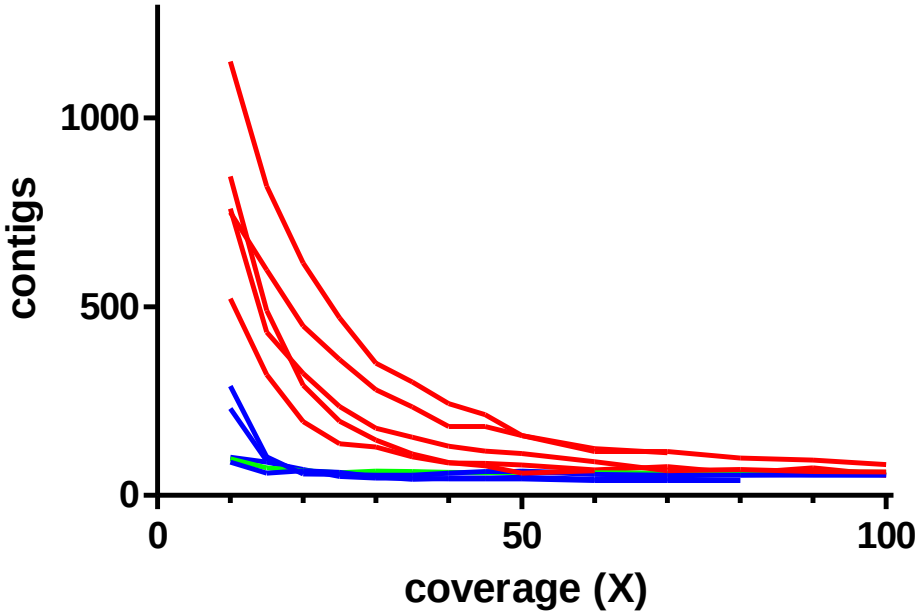
Library prep kit and sequencing instrument

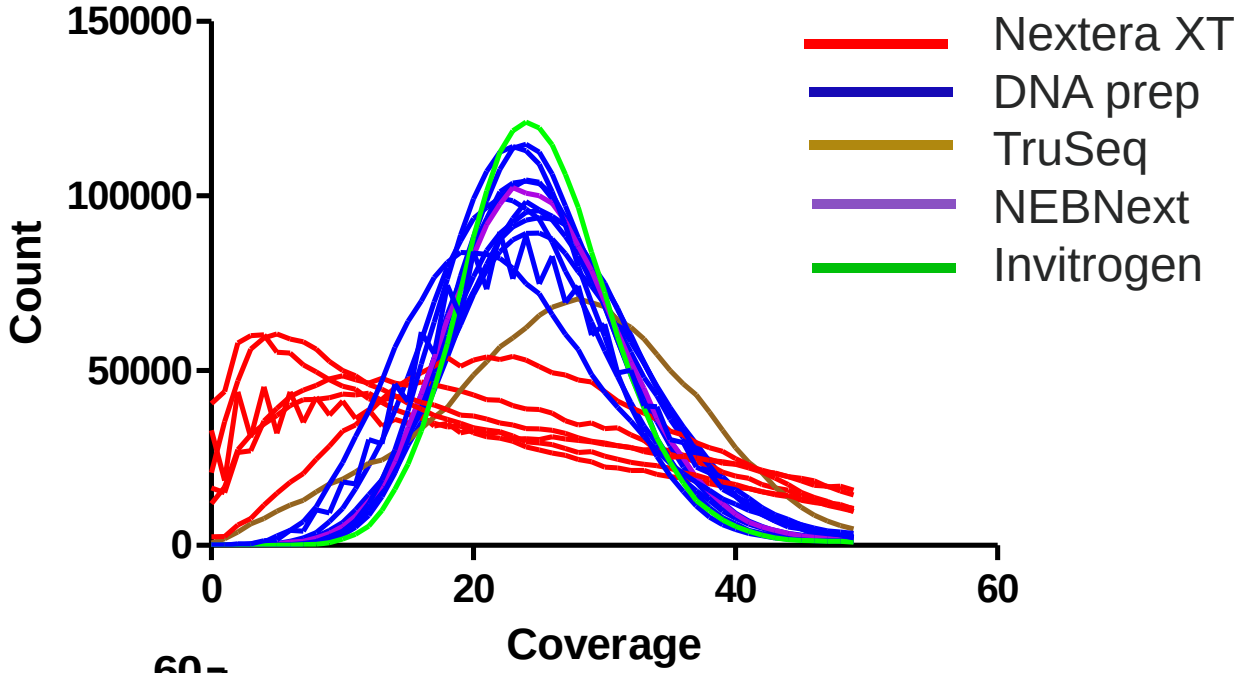


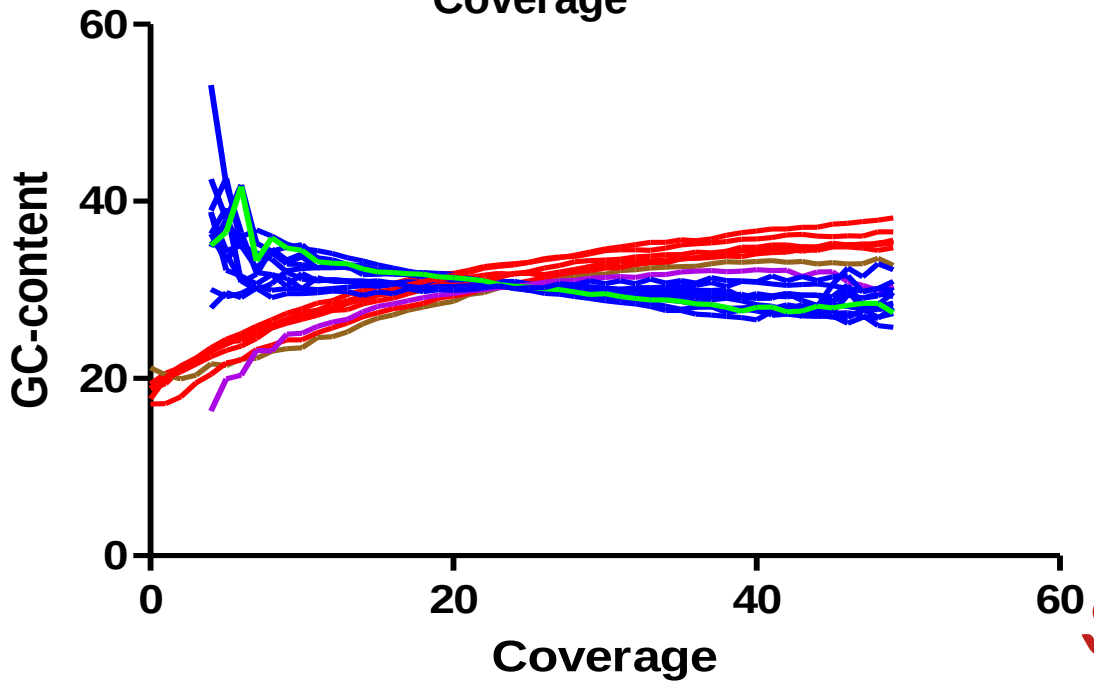


Library prep:
Nextera XT – Yes or No

- Nextera XT libraries result in an uneven distribution of the reads over the genome
- This bias is GC-content dependent
  - Low GC content regions have low coverage
  - High GC content regions high coverage

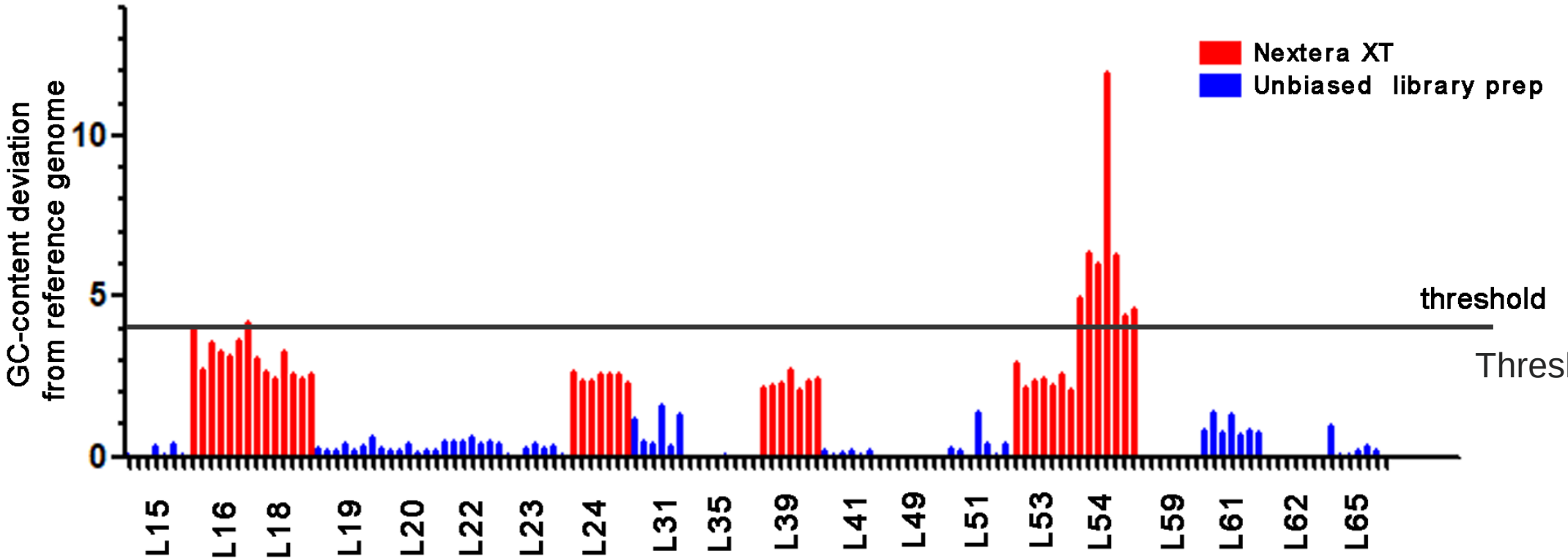






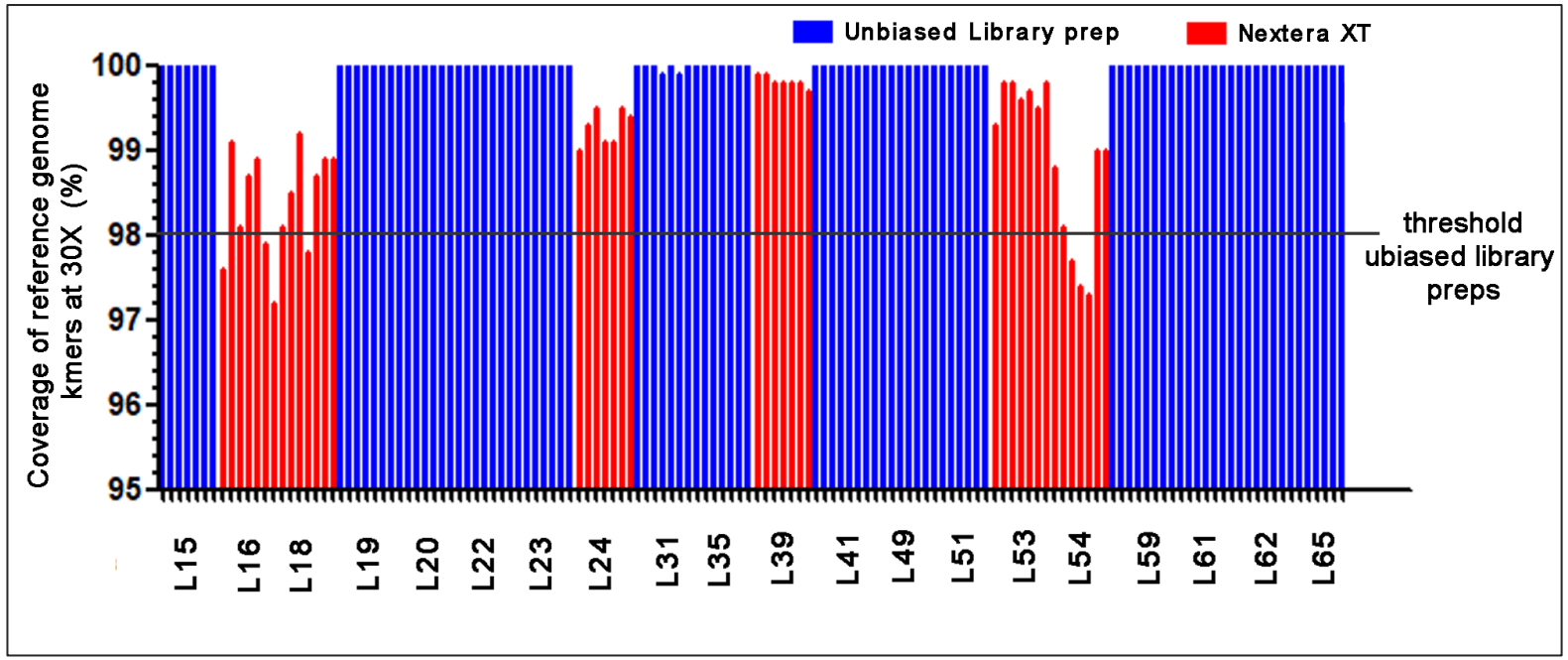
Library prep:
Nextera XT – Yes or No

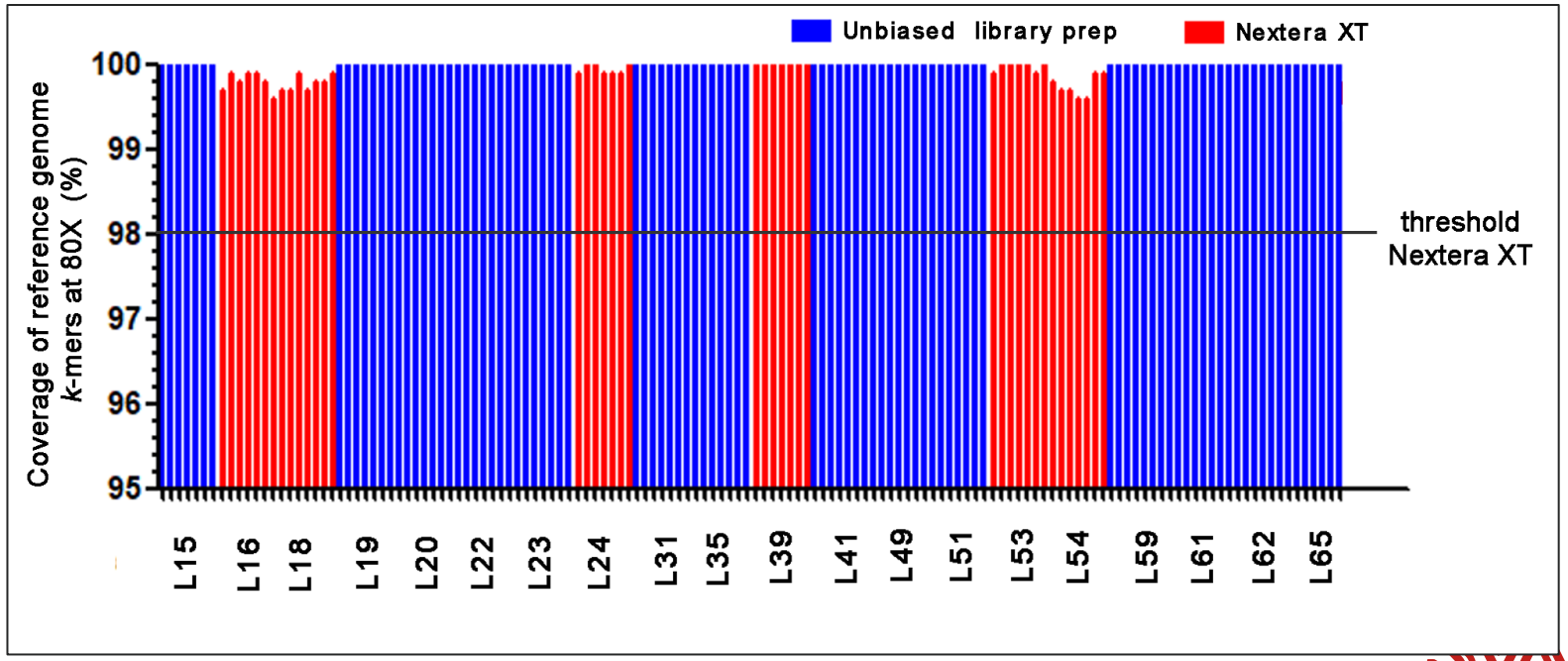
GC-content deviation in reads compared to reference genome



Threshold from ISO

Library prep:
Nextera XT – Yes or No





frontiers | Frontiers in Microbiology

TYPE Original Research
PUBLISHED 14 July 2022
DOI 10.3389/fmicb.2022.944770

Check for updates

OPEN ACCESS

EDITED BY
Ben Pascoe,
University of Bath, United Kingdom

REVIEWED BY
Abdul Karim Sesay,
Medical Research Council The Gambia Unit (MRC), Gambia
Craig T. Parker,
Agricultural Research Service,
United States
D. J. Darwin Bandoy,
University of the Philippines Los Baños,
Philippines

*CORRESPONDENCE
Bo Segerman
bo.segerman@sva.se

SPECIALTY SECTION
This article was submitted to Evolutionary and Genomic Microbiology, a section of the journal Frontiers in Microbiology

RECEIVED 16 May 2022
ACCEPTED 28 June 2022
PUBLISHED 14 July 2022

CITATION
Segerman B, Ástvaldsson Á, Mustafa L, Skarin J and Skarin H (2022) The efficiency of Nextera XT tagmentation depends on G and C bases in the binding motif leading to uneven coverage in bacterial species with low and neutral GC-content. *Front. Microbiol.* 13:944770. doi: 10.3389/fmicb.2022.944770

COPYRIGHT
© 2022 Segerman, Ástvaldsson, Mustafa, Skarin and Skarin. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

The efficiency of Nextera XT tagmentation depends on G and C bases in the binding motif leading to uneven coverage in bacterial species with low and neutral GC-content

Bo Segerman[1,2]*, Ásgeir Ástvaldsson[1], Linda Mustafa[2], Joakim Skarin[1,3] and Hanna Skarin[1]

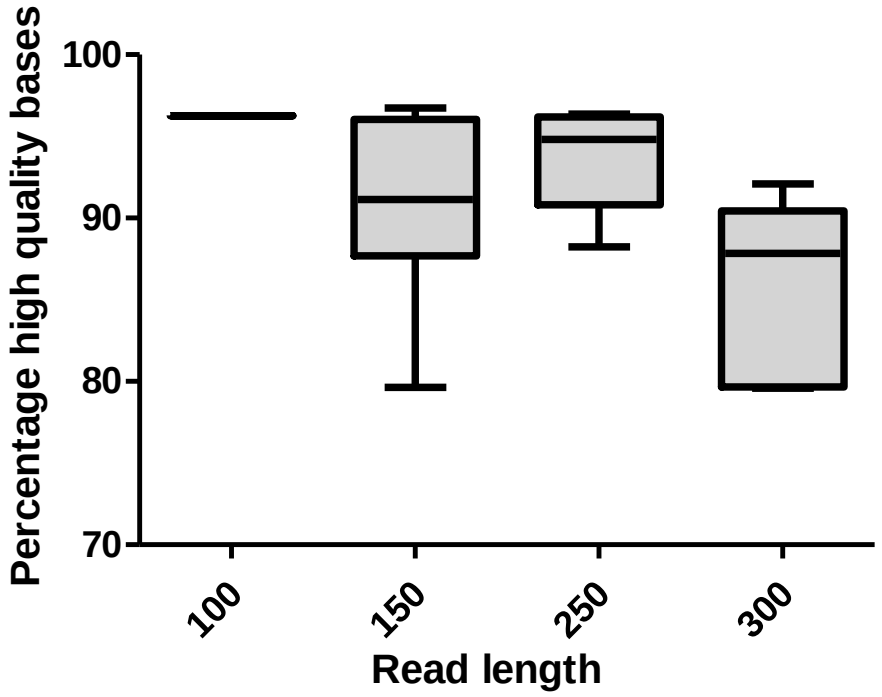
[1]Department of Microbiology, National Veterinary Institute (SVA), Uppsala, Sweden, [2]Department of Medical Biochemistry and Microbiology, Uppsala University, Uppsala, Sweden, [3]Department of Biology, Swedish Food Agency, Uppsala, Sweden

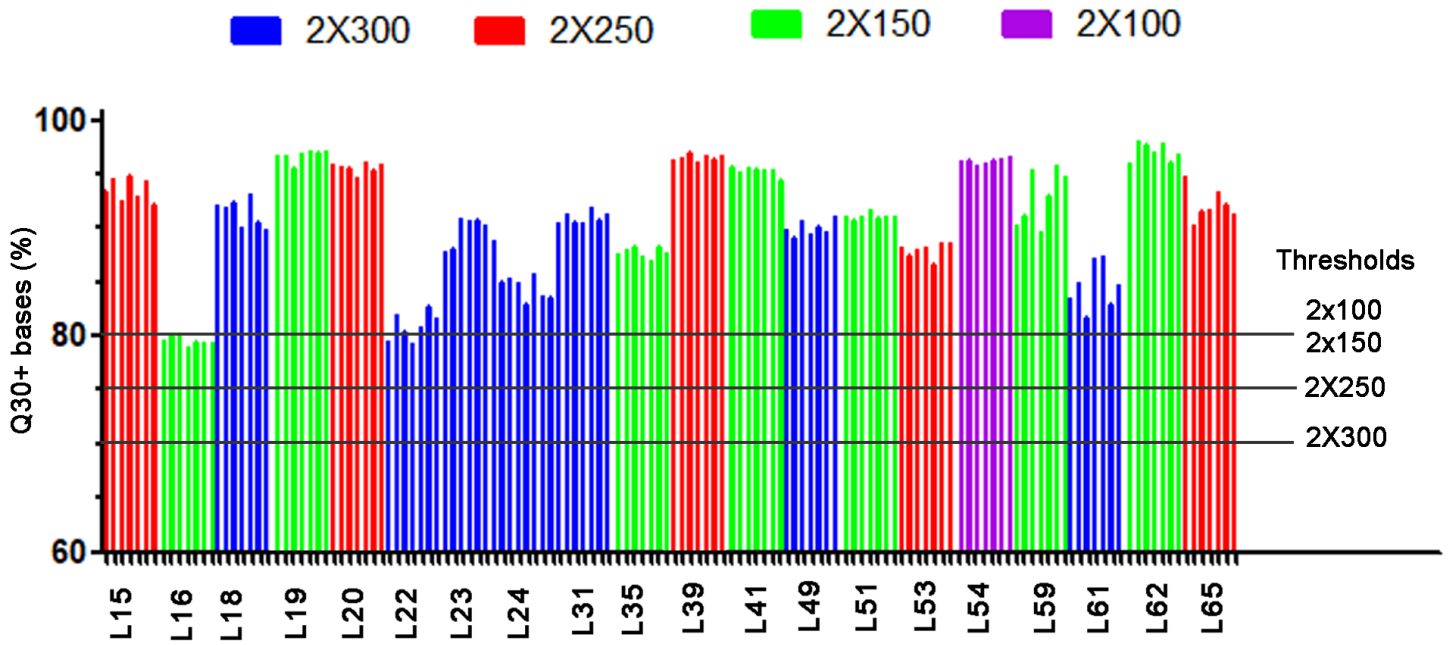
Whole-genome sequencing (WGS) is becoming the new standard for bacterial high-resolution typing and the performance of laboratories is being evaluated in interlaboratory comparisons. The use of the Illumina Nextera XT library preparation kit has been found to be associated with poorer performance due to a GC-content-dependent coverage bias. The bias is especially strong when sequencing low GC-content species. Here, we have made an in-depth analysis of the Nextera XT coverage bias problem using data from a proficiency test of the low GC-content species *Campylobacter jejuni*. We have compared Nextera XT with Nextera Flex/DNA Prep and examined the consequences on downstream WGS analysis when using different quantities of raw data. We have also analyzed how the coverage bias relates to differential usage of tagmentation cleavage sites. We found that the tagmentation site was characterized by a symmetrical motif with a central AT-rich region surrounded by Gs and Cs. The Gs and Cs appeared to be the main determinant for cleavage efficiency and the genomic regions that were associated with low coverage only contained low-efficiency cleavage sites. This explains why low GC-content genomes and regions are more subjected to coverage bias. We furthermore extended our analysis to other datasets representing other bacterial species. We visualized how the coverage bias was large in low GC-content species such as *C. jejuni*, *C. coli*, *Staphylococcus aureus*, and *Listeria monocytogenes*, whereas species with neutral GC-content such as *Salmonella enterica* and *Escherichia coli* were only affected in certain regions. Species with high GC-content such as *Mycobacterium tuberculosis* and *Pseudomonas aeruginosa* were hardly affected at all. The coverage bias associated with Nextera XT was not found when Nextera Flex/DNA Prep had been used.

KEYWORDS
Nextera XT, uneven, coverage, GC, bacterial, genome, *Campylobacter*

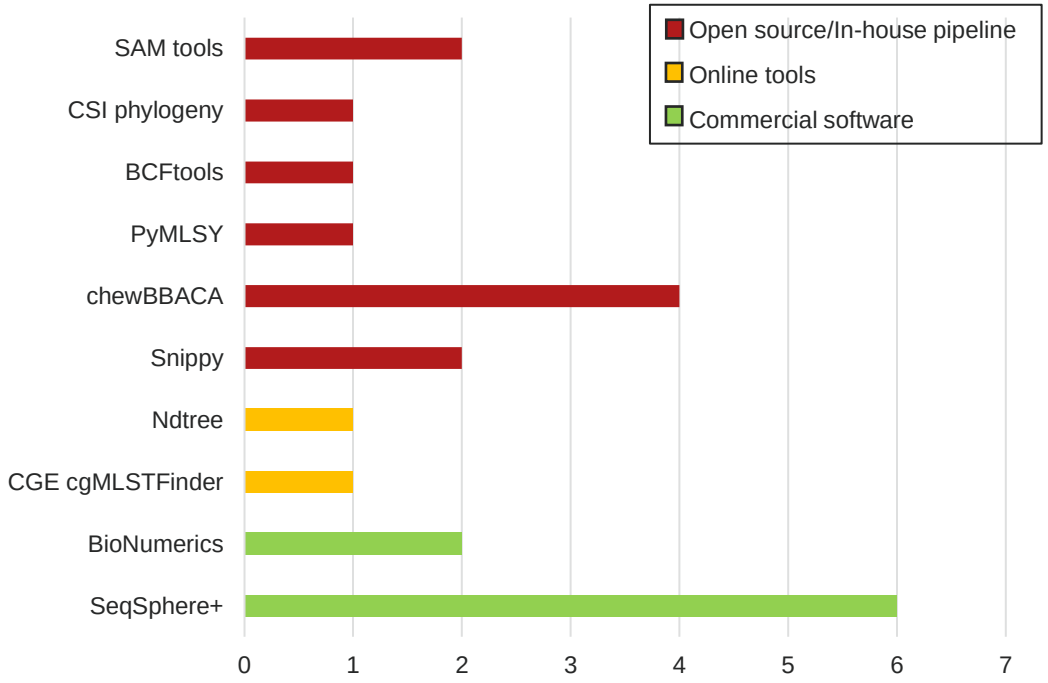
Read length:
Long or short read length

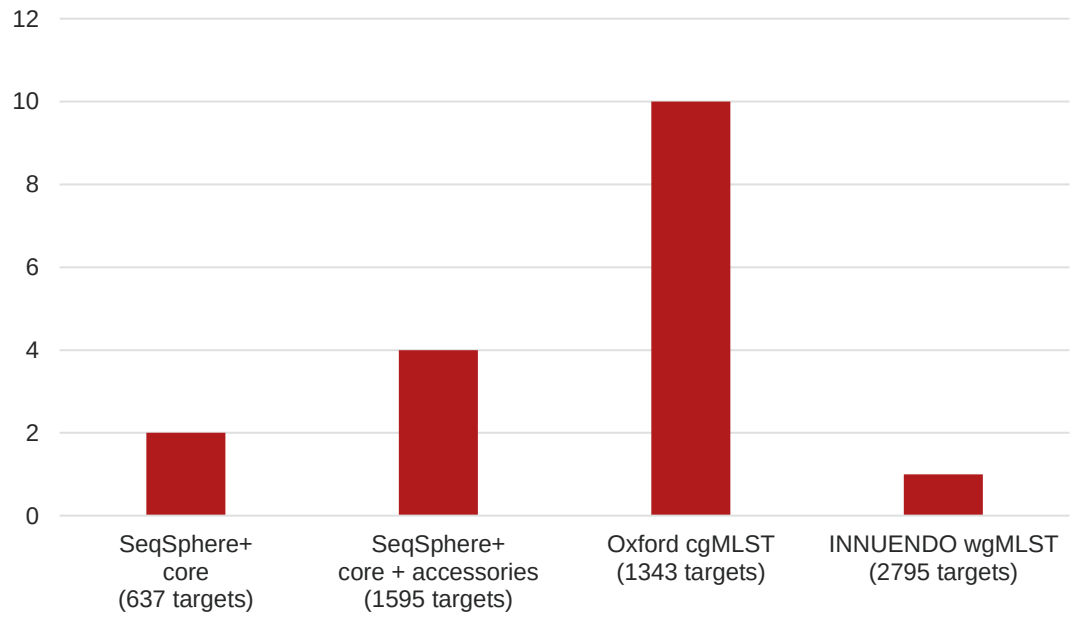
300 bp read length is associated with a noticeable quality drop





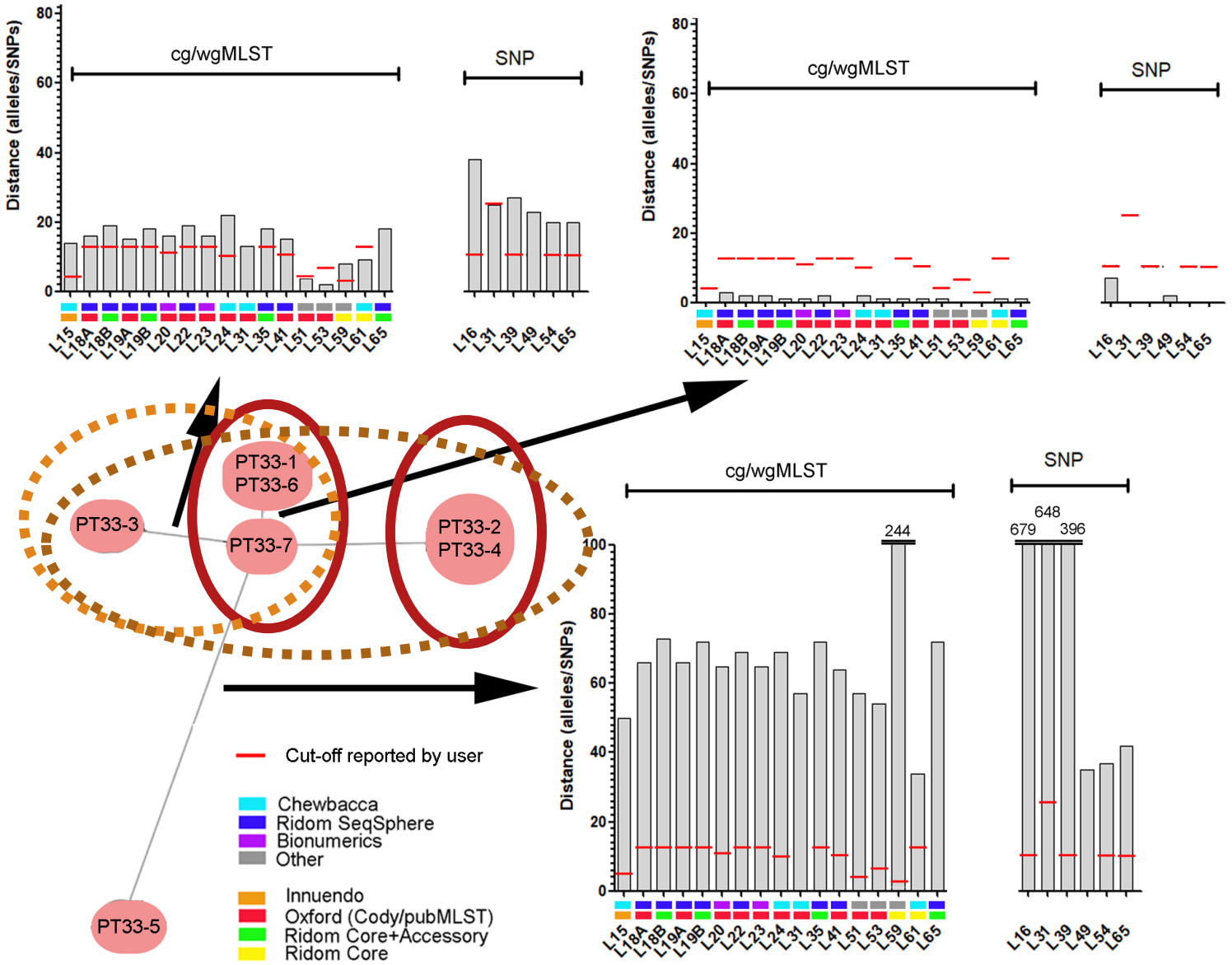
Cluster analysis





Cluster analysis

- The cluster cut-off values vary between NRLs.
- Still, most NRLs divide the samples into the same cluster structure.



Summary:

Nextera XT - Yes or no

requires higher coverage

gives GC deviation

Is it cheaper to use?

Read length 300 bp gives large quality drop compared to 250 bp

the last 50 bp are almost all trimmed off

consider using 250 read length > not as much data, but higher quality

Clustering interpretation is affected by method, software solution, schema, cut-off values used

Does not affect topology of the cluster analysis

Ridom SeqSphere+ core genome (cgMLST) schema is small and requires lower cut-off values

Ridom SeqSphere+, Chewbacca and Bionumerics perform similar.

Inter-EURLs working group on NGS – Reference WGS collection

Datasets from EURL-*Campylobacter* PT28 and PT33 can be obtained from the reference WGS collection. These datasets are very useful for validation and benchmarking.

Available through our website: https://www.sva.se/en/about-us/eurl-campylobacter/laboratory-procedures/inter-eurls-working-group-on-next-generation-sequencing/

WGS activities

- **EURL-*Campylobacter***
  - Training course on the analysis of WGS data from *Campylobacter*, requires some basic skills, January 25th-26th 2024, SVA, Sweden
  - Proficiency test number 38, 2024
    - Sequence quality and cluster analysis
    - Larger ”insilico” dataset and two DNA sample
- **Inter-EURLs WG on NGS**
  - Joint EURLs trainings course on NGS, basic level, June 2023, Netherlands
  - 2023 online webinar on organising PT-WGS

EUROPEAN STANDARD

NORME EUROPÉENNE

EUROPÄISCHE NORM

EN ISO 23418

June 2022

ICS 07.100.30

English Version

Microbiology of the food chain - Whole genome sequencing for typing and genomic characterization of bacteria - General requirements and guidance (ISO 23418:2022)

Microbiologie de la chaîne alimentaire - Séquençage de génome entier pour le typage et la caractérisation génomique des bactéries - Exigences générales et recommandations (ISO 23418:2022)

Mikrobiologie der Lebensmittelkette - Vollständige Genomsequenzierung zur Typisierung und genomischen Charakterisierung von Bakterien in Lebensmitteln - Allgemeine Anforderungen und Leitfaden (ISO 23418:2022)

This European Standard was approved by CEN on 20 May 2022.

CEN members are bound to comply with the CEN/CENELEC Internal Regulations which stipulate the conditions for giving this European Standard the status of a national standard without any alteration. Up-to-date lists and bibliographical references concerning such national standards may be obtained on application to the CEN-CENELEC Management Centre or to any CEN member.

This European Standard exists in three official versions (English, French, German). A version in any other language made by translation under the responsibility of a CEN member into its own language and notified to the CEN-CENELEC Management Centre has the same status as the official versions.

CEN members are the national standards bodies of Austria, Belgium, Bulgaria, Croatia, Cyprus, Czech Republic, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Italy, Latvia, Lithuania, Luxembourg, Malta, Netherlands, Norway, Poland, Portugal, Republic of North Macedonia, Romania, Serbia, Slovakia, Slovenia, Spain, Sweden, Switzerland, Turkey and United Kingdom.

EUROPEAN COMMITTEE FOR STANDARDIZATION
COMITÉ EUROPÉEN DE NORMALISATION
EUROPÄISCHES KOMITEE FÜR NORMUNG

CEN-CENELEC Management Centre: Avenue Marnix 17, B-1000 Brussels

© 2022 CEN All rights of exploitation in any form and by any means reserved worldwide for CEN national Members

Ref. No. EN ISO 23418:2022: E

Foreword

The WG has been established by the European Commission with the aim to promote the use of NGS across the EURLs' networks, build NGS capacity within the EU and ensure liaison with the work of the EURLs and the work of EFSA and ECDC on the NGS mandate sent by the Commission. The WG includes all the EURLs operating in the field of the microbiological contamination of food and feed and this document represents a deliverable of the WG and is meant to be diffused to all the respective networks of NRLs.

Guidance document for cluster analysis of whole genome sequence data

Version 02

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 DG-SANTE. Neither the European Union nor DG-SANTE can be held responsible for them.

Team EURL-*Campylobacter*

Thank you for your attention!

Questions?

Funded by
the European Union