Proficiency test number 33

WGS and Cluster Analysis of *Campylobacter*

Ásgeir Ástvaldsson
Bo Segerman

EURL-*Campylobacter* workshop
Sigtuna, September 27th, 2022

Co-funded by the European Union

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 join capability of the network to solve a multi-country *Campylobacter* outbreak based on WGS data

Participation

- 23 NRLs registered for PT 33
  - 18 EU member states
  - Norway, Switzerland and United Kingdom

- 20 NRLs reported results
  - 17 EU member states
  - Norway and United Kingdom

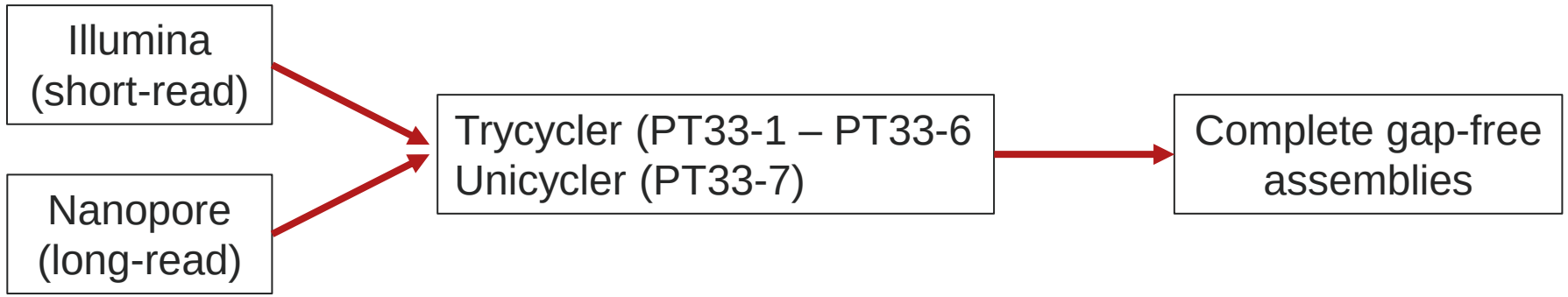
Strain selection

- Six strain of *Campylobacter jejuni* ST-19 selected based on cluster analysis topology
- Seven samples prepared for PT 33
  - O/N Cultures → DNA extraction → mixed with GenTegra-DNA → Aliquoted → Dried
- PT33-1 and PT33-6 the same sample
- Different timepoints
- Different farms

Sample	Strain	Matrix	Location	Sampling
PT33-1, PT33-6	20C120	Chicken caeca	Sweden, farm A	October, 2020
PT33-2	20C028	Chicken caeca	Sweden, farm B	July, 2020
PT33-3	20C126	Chicken caeca	Sweden, farm C	July, 2020
PT33-4	20C060	Chicken caeca	Sweden, farm A	July, 2020
PT33-5	Val_Cj015	Milk filter	Sweden, farm D	2011
PT33-7	20C102	Chicken caeca	Sweden, farm E	October, 2020

Reference assemblies

- Reference assemblies generated for all strains



Sample	No. of contigs	Assembly size (bp)	GC %	Assembly pipeline
PT33-1, PT33-6	1	1711096	30,47	Trycycler
PT33-2	1	1753524	30,40	Trycycler
PT33-3	1	1673246	30,48	Trycycler
PT33-4	1	1753518	30,40	Trycycler
PT33-5	1	1760653	30,34	Trycycler
PT33-7	1	1711097	30,48	Unicycler

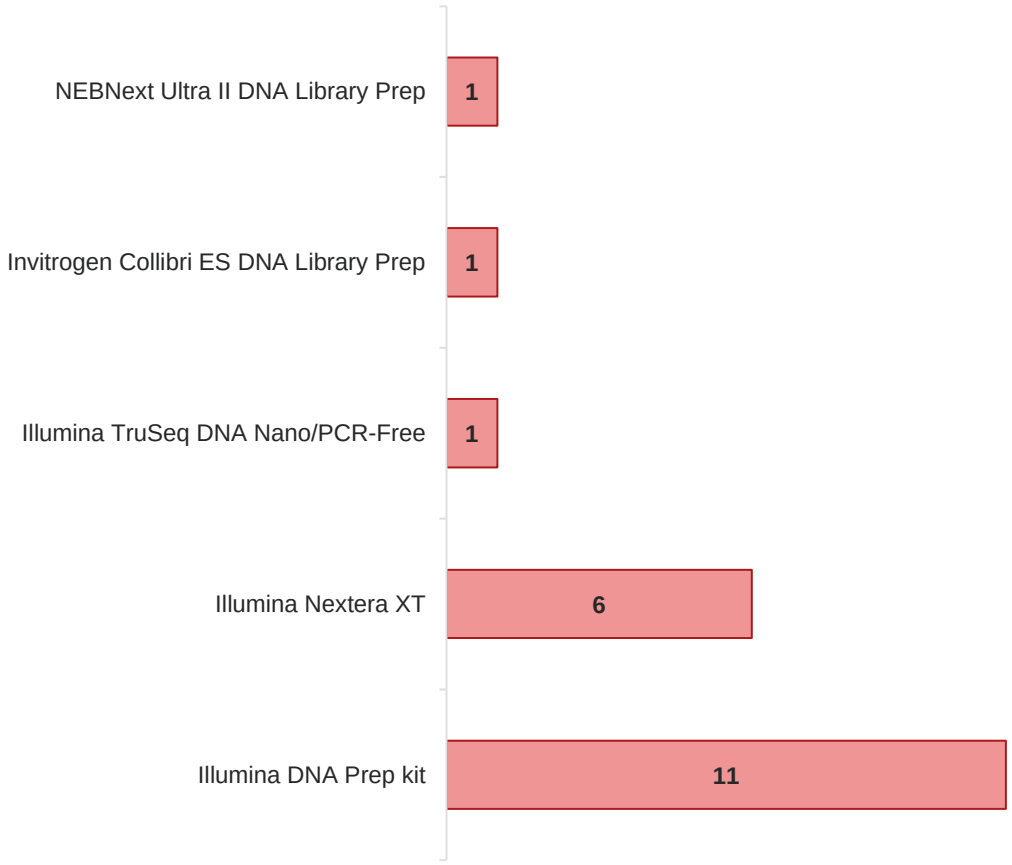
Outline

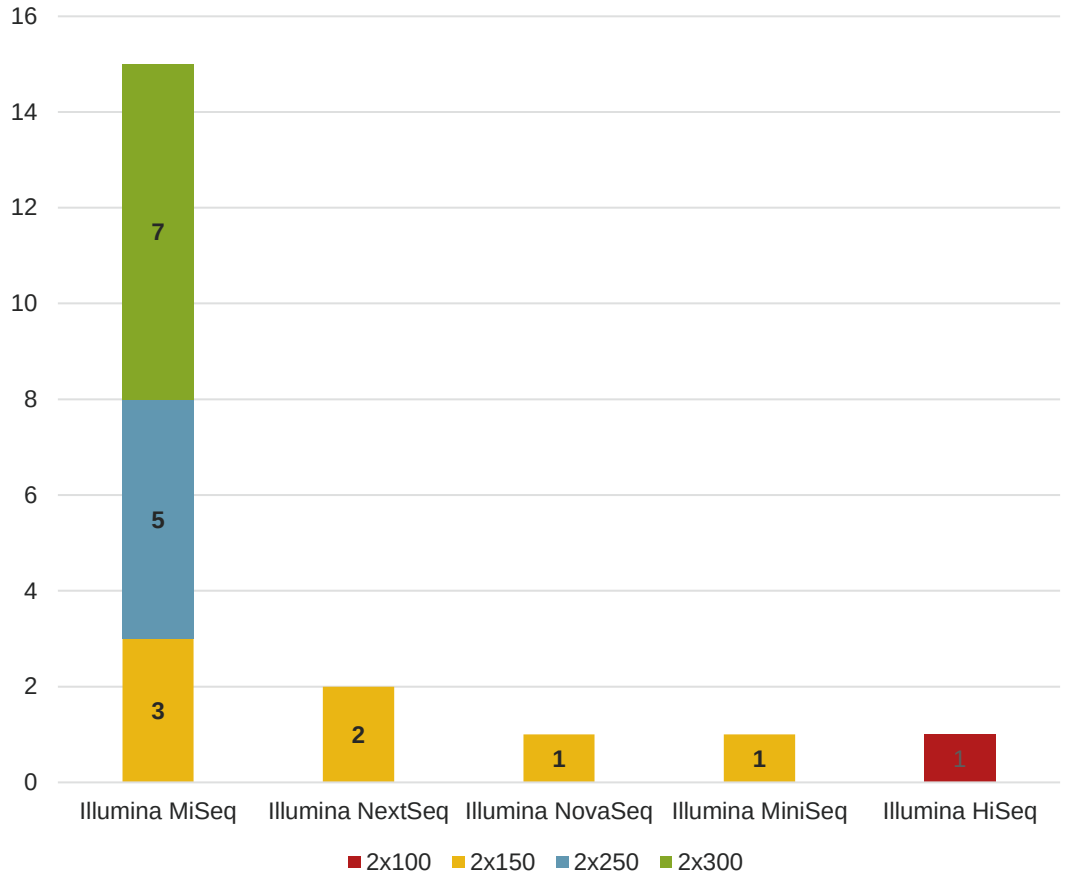
- Samples distributed together with the other PTs, 7th March 2022
- Process samples according to standard laboratory procedures
  - Library preparations → Sequencing → Downstream analysis
    - MLST analysis
    - AMR analysis (optional)
    - Cluster analysis
      - Gene-by-gene, SNP or other methods
      - Use own cut-off value for cluster analysis

Outline

- Last date for reporting of results: 1st of June, 2022
- Questback questionnaire
- Upload data to a personal OneDrive folder
  - Raw sequencing data (fastq)
  - Assemblies (if required for cluster analysis)
  - Trees for cluster analysis (e.g. phylogenetic or minimum spanning)
  - Raw clustering data (e.g. distance matrix or alignment)

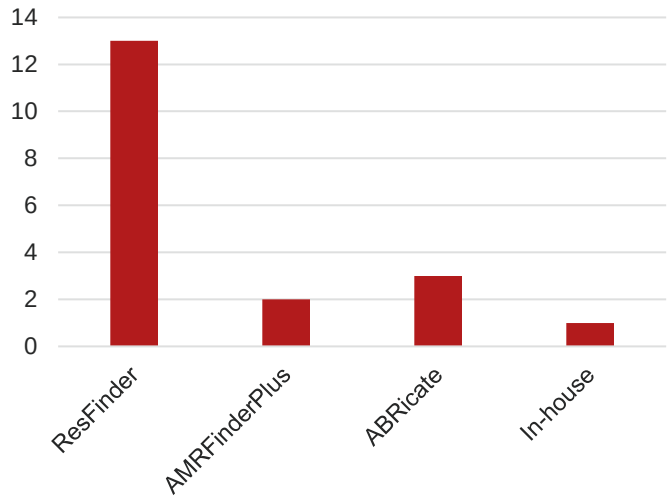
Library prep kit and sequencing instrument

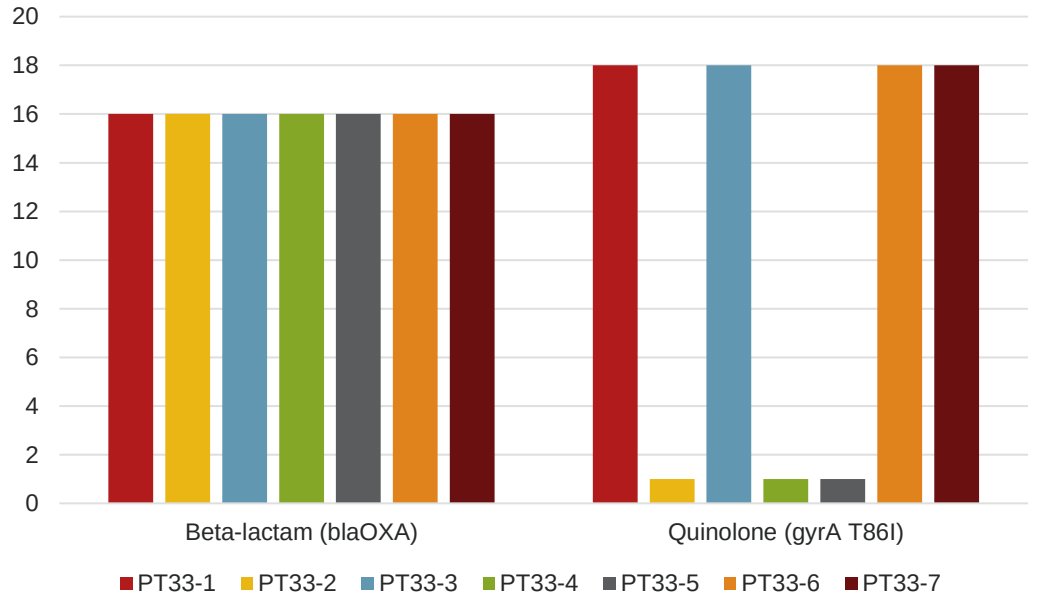




AMR analysis (optional)

19 NRLs reported results for the optional AMR analysis





- **Not used for assessment of the participating NRLs**

Assemblies

- Participants were asked to submit the assemblies, if part of analysis
- 17 NRLs submitted assemblies (16 used SPAdes, 1 used Velvet)
- Calculated QC metrics:

QC metric	Results	Median
Total size of assembly (bp)	<2% deviation from reference for all samples except PT33-4	
k-mer coverage over the reference genome (%)	>99% for all samples	99.978%
Total number of contigs	*17 – 68 contigs	*32 contigs
Total number of contigs > 1kb	*13 – 32 contigs	*18 contigs
Longest contig	*188,052 bp – 720,947	*440,801 bp
N50 length	*79,979 bp – 216,952 bp	*154,047 bp

PT33-4 and L20 excluded

- **Not used for assessment of the participating NRLs**

Performance assessment

- No overall performance grade
- Individual steps assessed → Satisfactory / Needs improvement

Criteria and cut-off values used for assessment of **sequence quality**

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.

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.

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

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

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.

frontiers | Frontiers in Microbiology

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

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

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

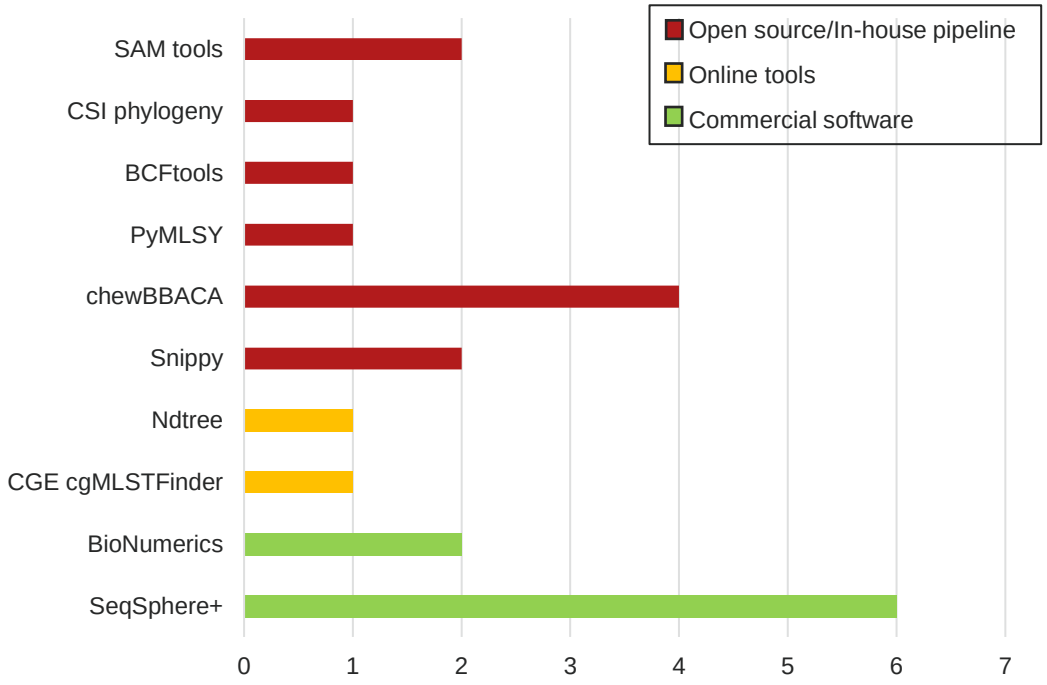
Whole-genome sequencing (WGS) is becoming the new standard for bacterial high-resolution typing and the performance of laboratories is being evaluated in interlaboratory comparisons. The use of the Illumina Nextera XT library preparation kit has been found to be associated with poorer performance due 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.

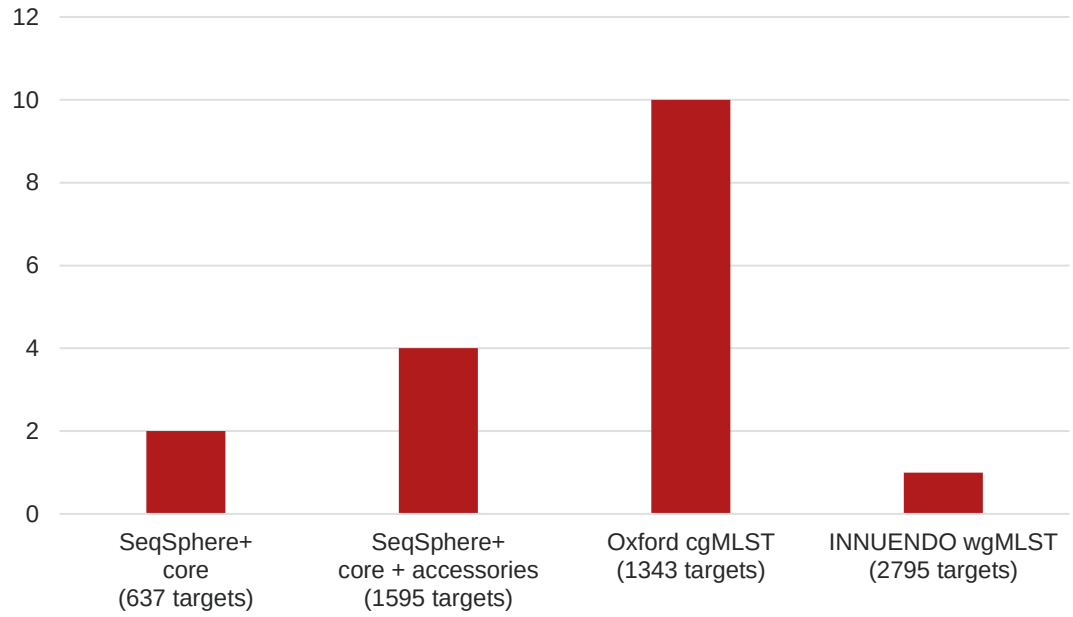
KEYWORDS
Nextera XT, uneven, coverage, GC, bacterial, genome, *Campylobacter*

Assessment of sequence quality

Lab code	MLST	Q30	Contamination	Reference coverage	GC deviation	Overall evaluation sequence quality
L15	6/6	6/6	6/6	6/6	6/6	Satisfactory
L16	6/6	2/6	6/6	6/6	5/6	Needs improvement
L18	6/6	6/6	6/6	6/6	6/6	Satisfactory
L19	6/6	6/6	6/6	6/6	6/6	Satisfactory
L20	6/6	6/6	6/6	6/6	6/6	Satisfactory
L22	6/6	6/6	6/6	6/6	6/6	Satisfactory
L23	6/6	6/6	6/6	6/6	6/6	Satisfactory
L24	6/6	6/6	6/6	6/6	6/6	Satisfactory
L31	6/6	6/6	6/6	6/6	6/6	Satisfactory
L35	6/6	6/6	6/6	6/6	6/6	Satisfactory
L39	6/6	6/6	6/6	6/6	6/6	Satisfactory
L41	6/6	6/6	6/6	6/6	6/6	Satisfactory
L49	6/6	6/6	6/6	6/6	6/6	Satisfactory
L51	6/6	6/6	6/6	6/6	6/6	Satisfactory
L53	6/6	6/6	6/6	6/6	6/6	Satisfactory
L54	6/6	6/6	6/6	6/6	0/6	Needs improvement
L59	6/6	6/6	6/6	6/6	6/6	Satisfactory
L61	6/6	6/6	6/6	6/6	6/6	Satisfactory
L62	6/6	6/6	6/6	6/6	6/6	Satisfactory
L65	6/6	6/6	6/6	6/6	6/6	Satisfactory

Cluster analysis





Assessment of cluster analysis

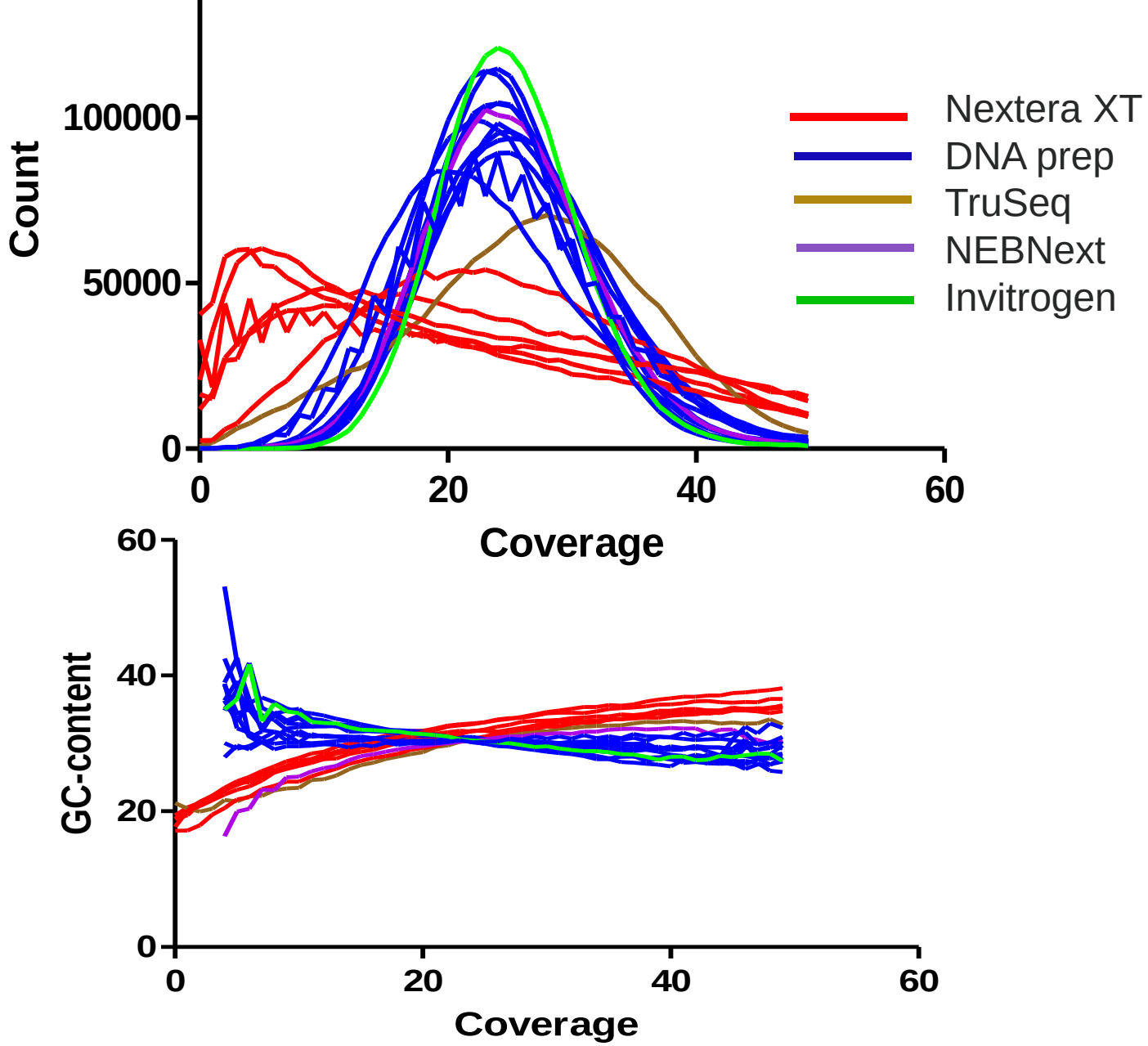
Lab code	PT33-6 and PT33-7 are the two closest samples to PT33-1	PT33-4 is the closest sample to PT33-2	PT33-5 is most distant to other samples	Overall evaluation sequence quality
15	+	+	+	Satisfactory
16	+	+	+	Satisfactory
18	+	+	+	Satisfactory
19	+	+	+	Satisfactory
20	+	+	+	Satisfactory
22	+	+	+	Satisfactory
23	+	+	+	Satisfactory
24	+	+	+	Satisfactory
31	+	+	+	Satisfactory
35	+	+	+	Satisfactory
39	+	+	+	Satisfactory
41	+	+	+	Satisfactory
49	+	+	+	Satisfactory
51	+	+	+	Satisfactory
53	+	+	+	Satisfactory
54	+	+	+	Satisfactory
59	+	+	+	Satisfactory
61	+	+	+	Satisfactory
62	+	+	+	Satisfactory
65	+	+	+	Satisfactory

Analysis of PT33 sequence data

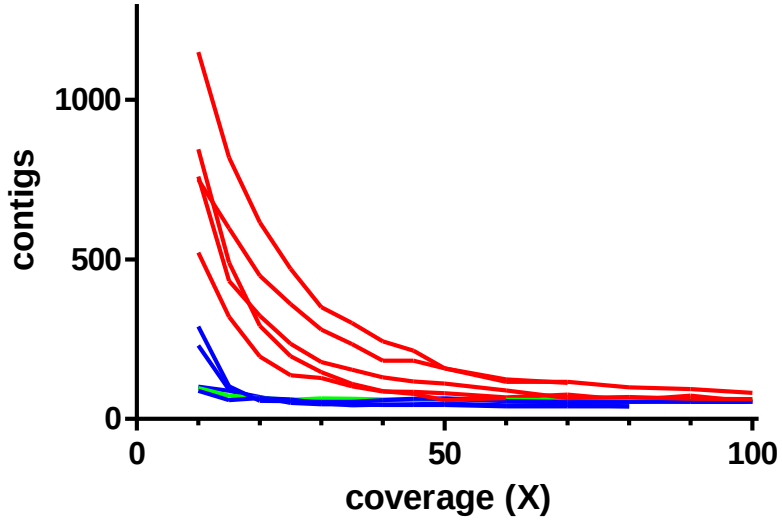
Bo Segerman

Library prep:
Nextera XT – Yes or No

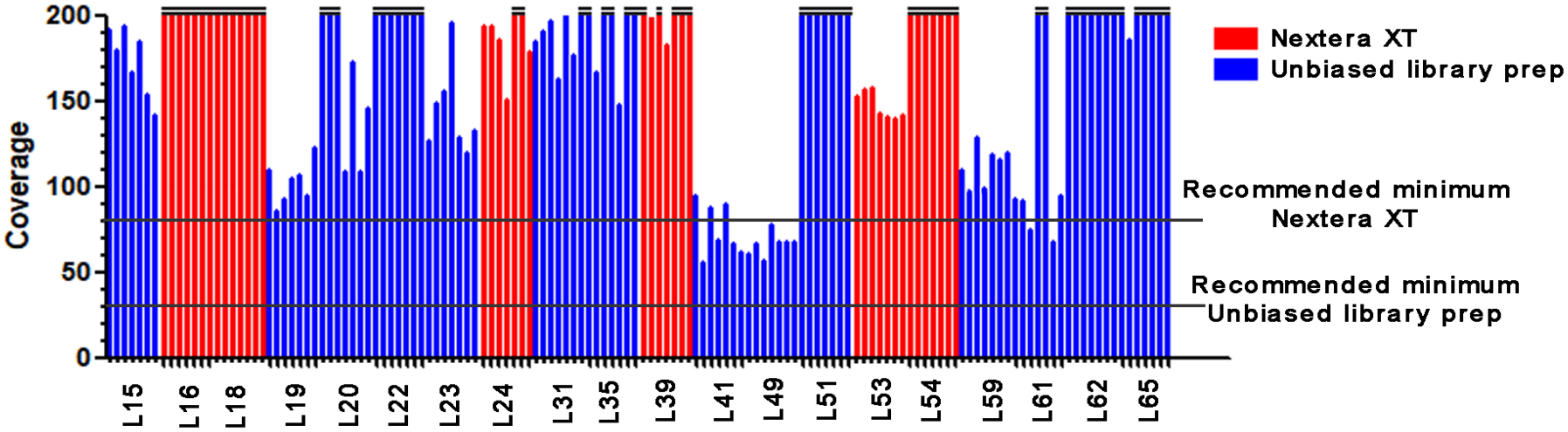
- Nextera XT has a uneven distribution of the reads over the genome
- This bias is GC-content dependent
  - Low GC content regions have low coverage
  – high GC content regions high coverage



Library prep:
Nextera XT – Yes or No

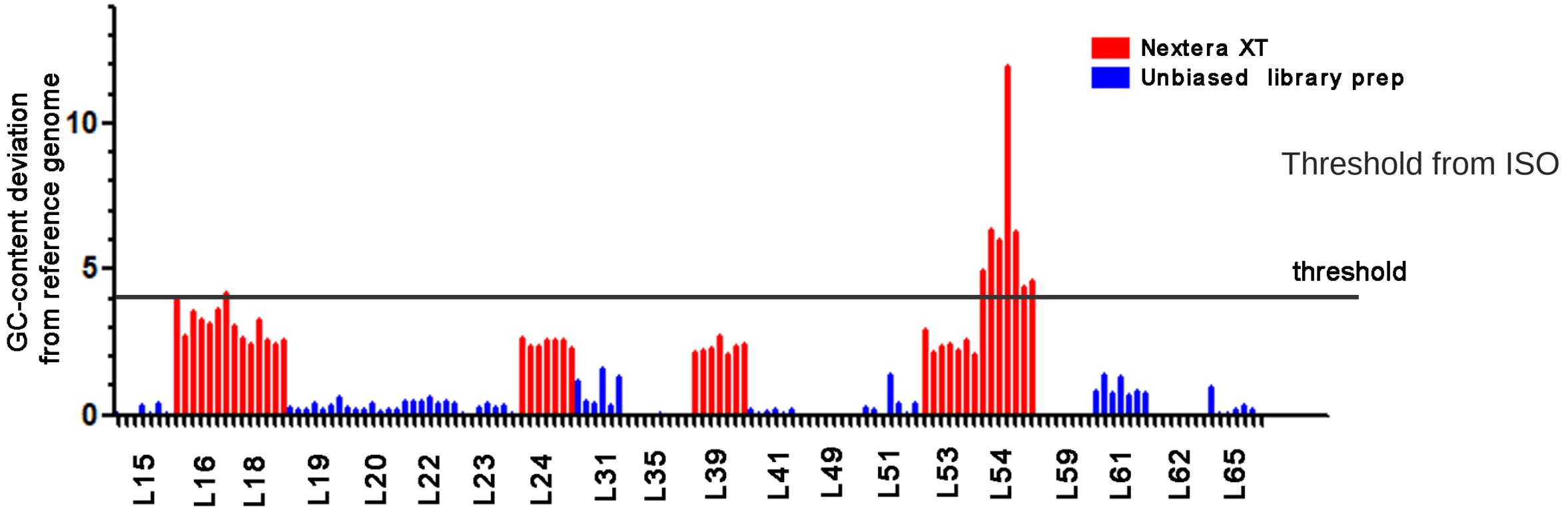


Different ”recommended” minimum coverage
Nextera XT or other library prep



Library prep:
Nextera XT – Yes or No

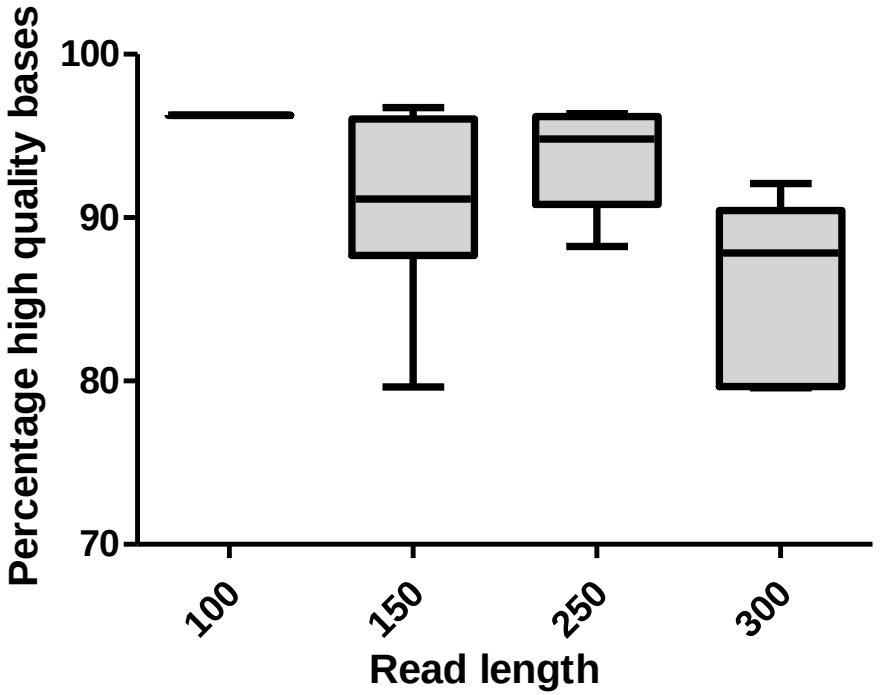
GC-content deviation in reads compared to reference genome



Threshold from ISO

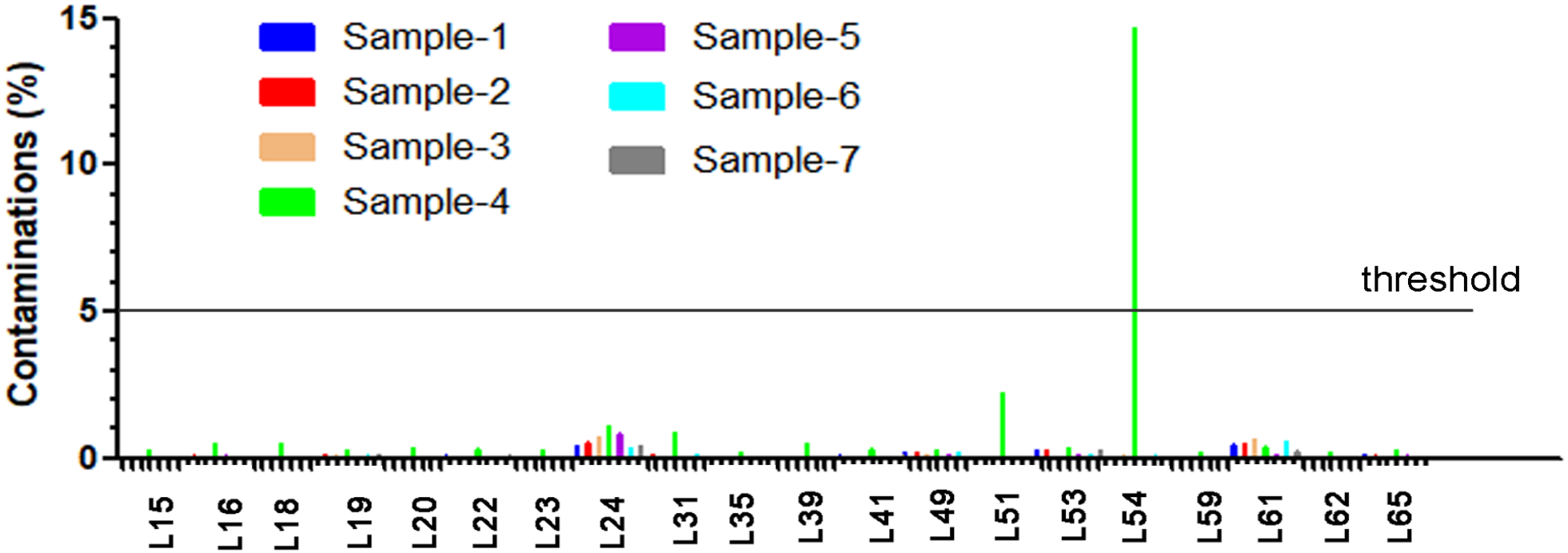
Read length:
Long or short read length

300 bp read length is associated with a noticeable quality drop



Contaminations:
What are the consequences?

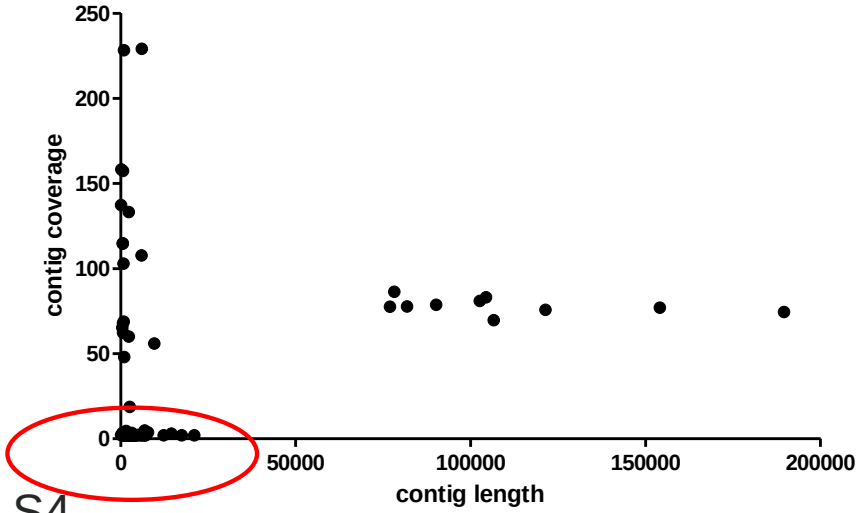
Contaminations are < 1% except sample 4



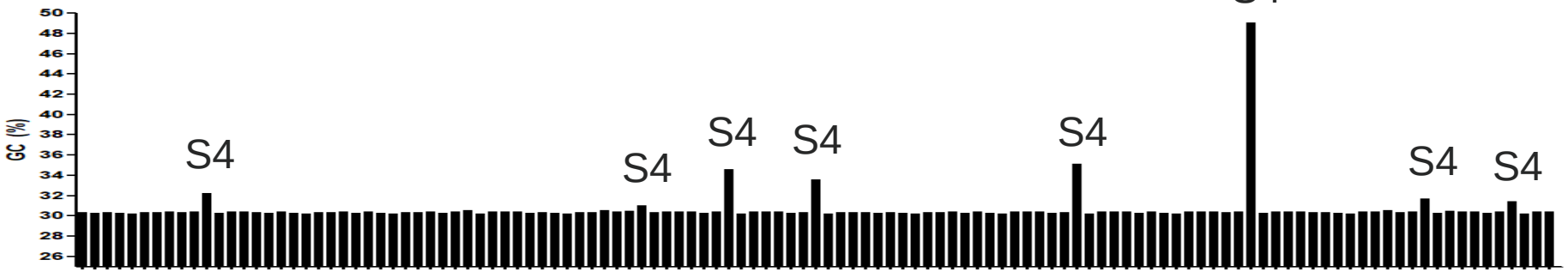
Contaminations:
What are the consequences?

Contaminations infiltrates the assemblies but can be filtered out
- if from an unrelated species

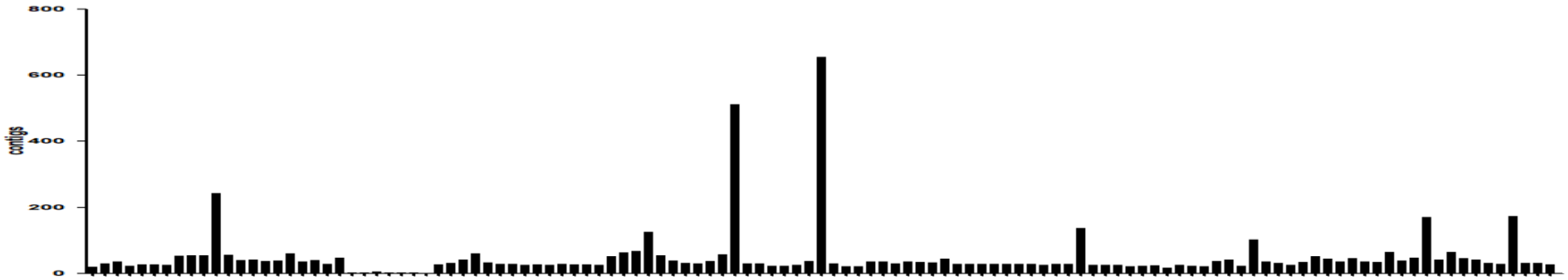
Unrelated contamination
- Little effect on clustering



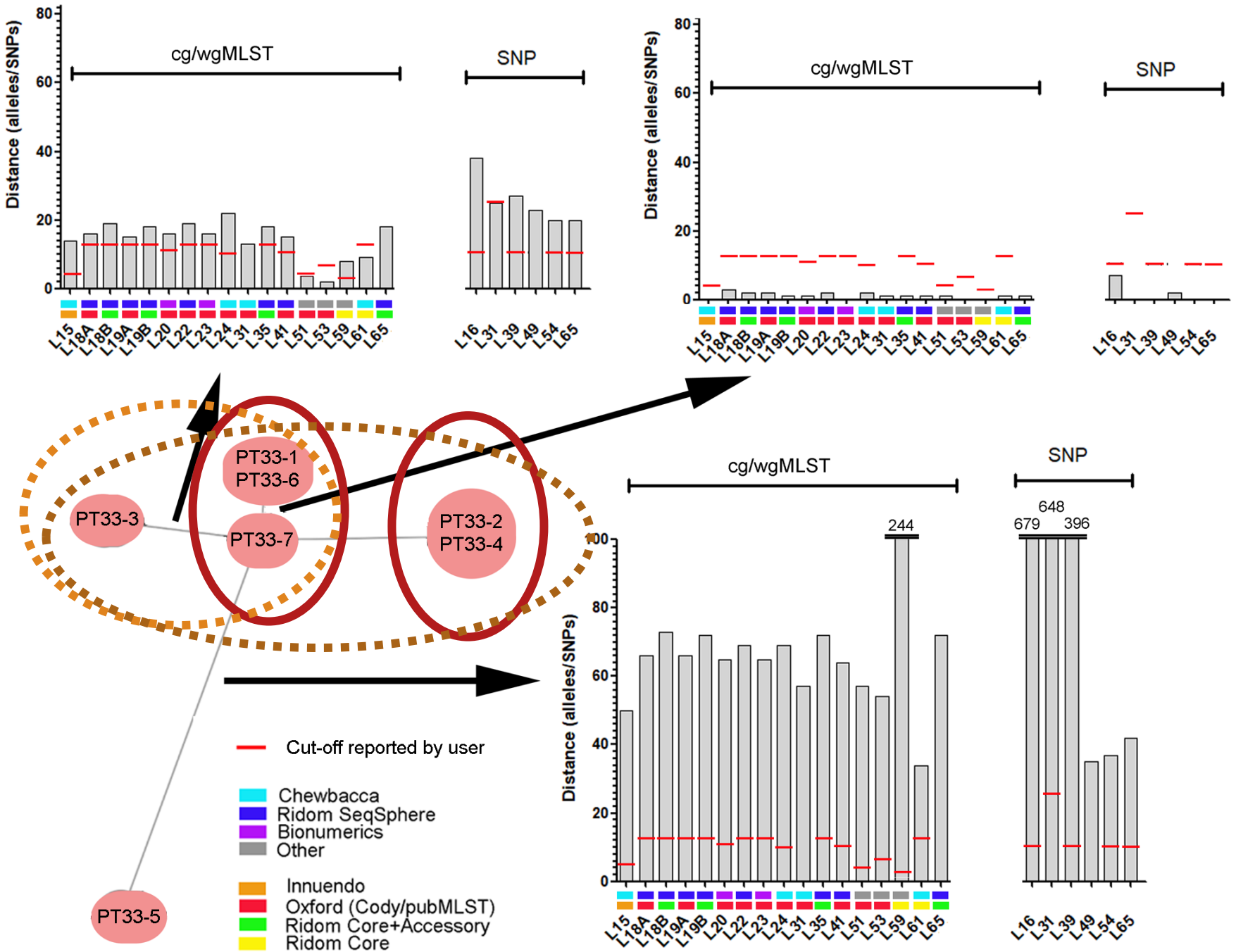
GC-content



Contig count



Clustering: Which factors influences the results?



Summary:

Overall satisfactory performance in the PT

Nextera XT - requires higher coverage
- gives GC deviation

Read length 300 gives large quality drop compared to 250

Contamination levels low (except one sample – uneven spread between samples)
- Contaminations propagate to the assemblies (poorer assembly metrics)
- Contamination of unrelated species can be filtered
- Contamination of unrelated species do not interfere with reference/schema based clustering methods

Clustering interpretation is affected by method, software solution, schema, cut-off values used
-Ridom SeqSphere+ core genome (cgMLST) schema is small and requires lower cut-off values
- Ridom SeqSphere+, Chewbacca and Bionumerics perform similar.