E. coli genomics and quality control (QC)

GenEpi-BioTrain Virtual training 23

Virulence Profiling of *E. coli* Pathotypes

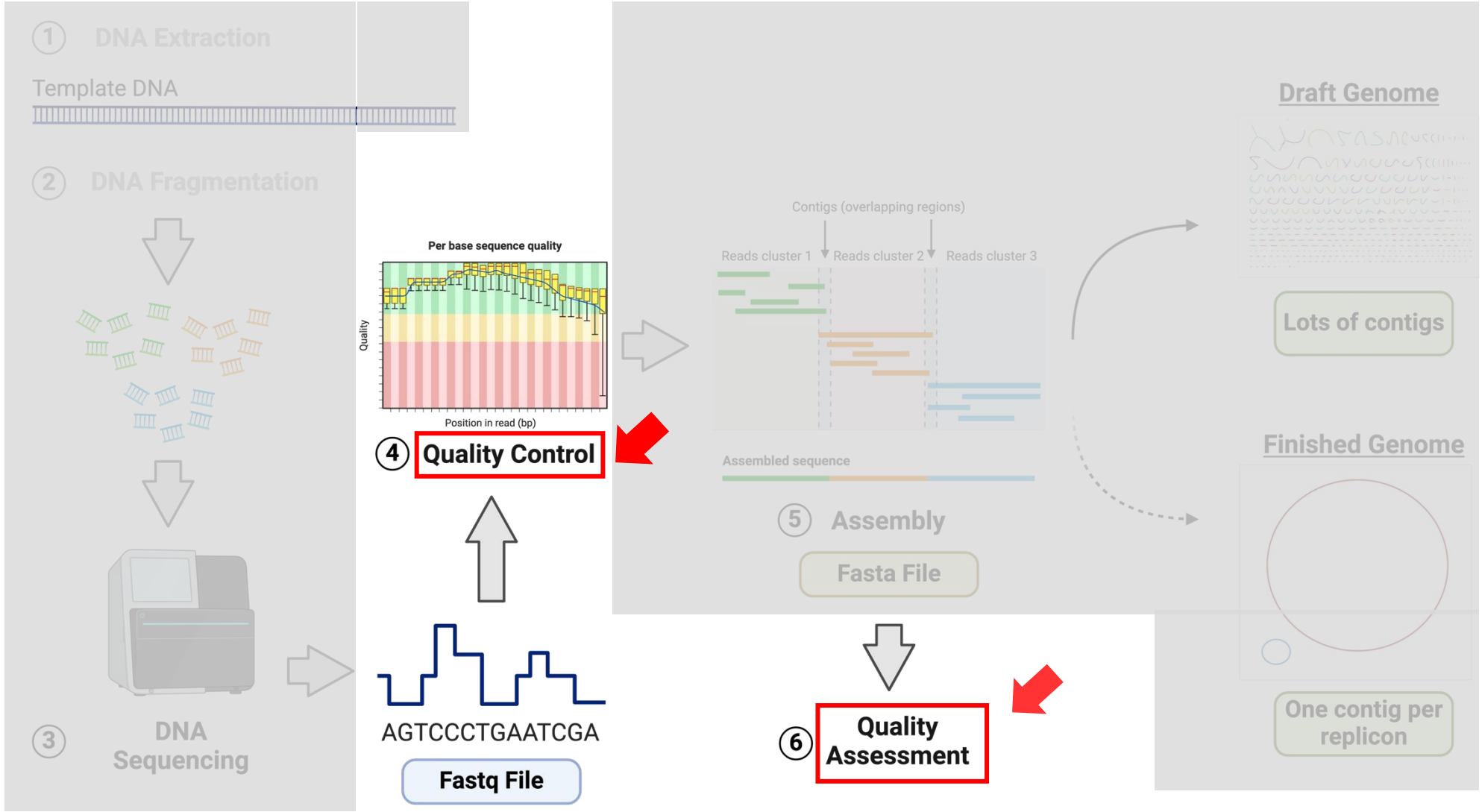
João Cardoso (joacar@dtu.dk)
Technical University of Denmark (DTU), National Food Institute

03/03/2026

Intended Learning Objectives 

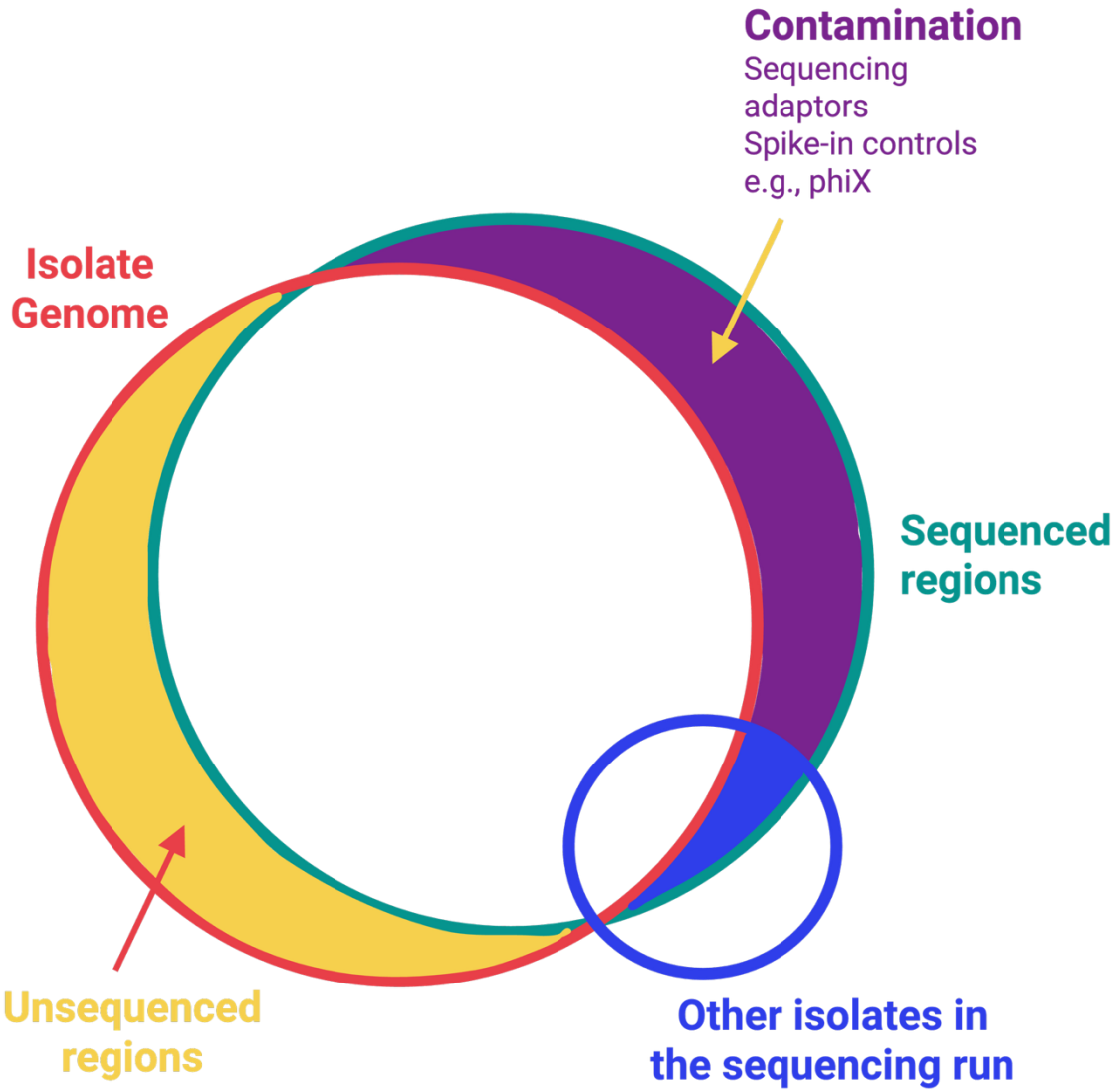
1. What is the importance of Quality Control (QC)?
2. How can we access the quality of the sequencing data?
3. What web-based tools are available?
4. What metrics should we evaluate?

Simplified WGS Analysis Pipeline



Adapted from Joana Mourão (joam@dtu.dk); Image Created with BioRender.com

What Data do We Really Have?



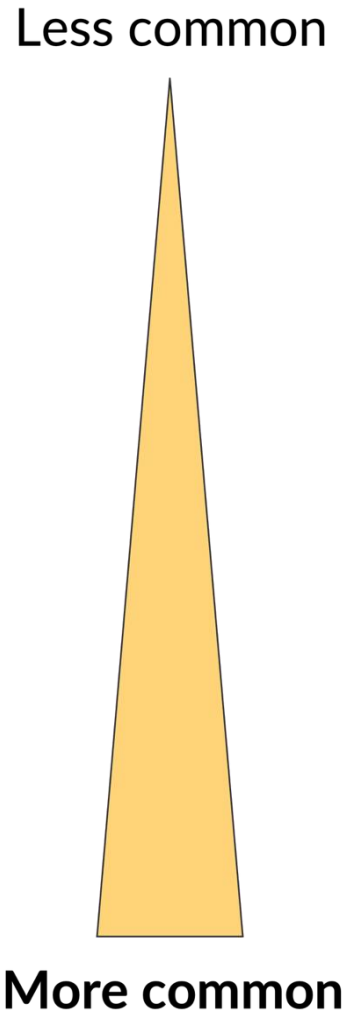
Adapted from Joana Mourão (joam@dtu.dk); Image Created with BioRender.com

Assessment of Quality of Sequencing Raw Data

FASTQC & FASTP

Sequencing Can Introduce Errors

Type of Error	Cause	Explanation
Nonsense reads	Instrument oddness	Calibration error, equipment malfunction
Duplicate reads	Amplify a low-complexity library	Same DNA fragment is sequenced multiple times – not enough DNA sample diversity
Adaptor read-through	Fragment too short	If too short bacterial DNA fragment – sequencing might read into adaptor sequence
Indel errors	Inserting extra bases, skipping/deleting bases	Could lead to frameshift mutations if errors occur within the coding sequences
Uncalled base	Couldn't reliably estimate, replace with "N"	'N' is used to represent an uncalled base in the genome
Substitution errors	Reading wrong base (e.g., read "A" instead of "G")	Sequencer incorrectly identifies a base – lead to single nucleotide polymorphism (SNP)



What is in my Fastq File?

FASTQ Format: A "standard" format for **storing** and **defining** sequences from NGS platforms

Each quality value is encoded by the **ASCII characters** – read by computers

Encoded quality values, one symbol per nucleotide

Calculated using various **parameters** – e.g., peak shape, peak resolution in each base, signal intensity

Quality Score Encoding - ASCII Table

Decimal	Hex	Char	Decimal	Hex	Char	Decimal	Hex	Char	Decimal	Hex	Char
0	0	[NULL]	32	20	[SPACE]	64	40	@	96	60	`
1	1	[START OF HEADING]	33	21	!	65	41	A	97	61	a
2	2	[START OF TEXT]	34	22	"	66	42	B	98	62	b
3	3	[END OF TEXT]	35	23	#	67	43	C	99	63	c
4	4	[END OF TRANSMISSION]	36	24	$	68	44	D	100	64	d
5	5	[ENQUIRY]	37	25	%	69	45	E	101	65	e
6	6	[ACKNOWLEDGE]	38	26	&	70	46	F	102	66	f
7	7	[BELL]	39	27	'	71	47	G	103	67	g
8	8	[BACKSPACE]	40	28	(	72	48	H	104	68	h
9	9	[HORIZONTAL TAB]	41	29	)	73	49	I	105	69	i
10	A	[LINE FEED]	42	2A	*	74	4A	J	106	6A	j
11	B	[VERTICAL TAB]	43	2B	+	75	4B	K	107	6B	k
12	C	[FORM FEED]	44	2C	,	76	4C	L	108	6C	l
13	D	[CARRIAGE RETURN]	45	2D	-	77	4D	M	109	6D	m
14	E	[SHIFT OUT]	46	2E	.	78	4E	N	110	6E	n
15	F	[SHIFT IN]	47	2F	/	79	4F	O	111	6F	o
16	10	[DATA LINK ESCAPE]	48	30	0	80	50	P	112	70	p
17	11	[DEVICE CONTROL 1]	49	31	1	81	51	Q	113	71	q
18	12	[DEVICE CONTROL 2]	50	32	2	82	52	R	114	72	r
19	13	[DEVICE CONTROL 3]	51	33	3	83	53	S	115	73	s
20	14	[DEVICE CONTROL 4]	52	34	4	84	54	T	116	74	t
21	15	[NEGATIVE ACKNOWLEDGE]	53	35	5	85	55	U	117	75	u
22	16	[SYNCHRONOUS IDLE]	54	36	6	86	56	V	118	76	v
23	17	[ENG OF TRANS. BLOCK]	55	37	7	87	57	W	119	77	w
24	18	[CANCEL]	56	38	8	88	58	X	120	78	x
25	19	[END OF MEDIUM]	57	39	9	89	59	Y	121	79	y
26	1A	[SUBSTITUTE]	58	3A	:	90	5A	Z	122	7A	z
27	1B	[ESCAPE]	59	3B	;	91	5B	[	123	7B	{
28	1C	[FILE SEPARATOR]	60	3C	<	92	5C	\	124	7C	|
29	1D	[GROUP SEPARATOR]	61	3D	=	93	5D	]	125	7D	}
30	1E	[RECORD SEPARATOR]	62	3E	>	94	5E	^	126	7E	~
31	1F	[UNIT SEPARATOR]	63	3F	?	95	5F	_	127	7F	[DEL]

The formula for getting the **PHRED score** from encoded quality:

Q = ascii(char) - 33

Example:

!+EJ

ASCII:	33	43	69	74
-33:	0	10	36	41

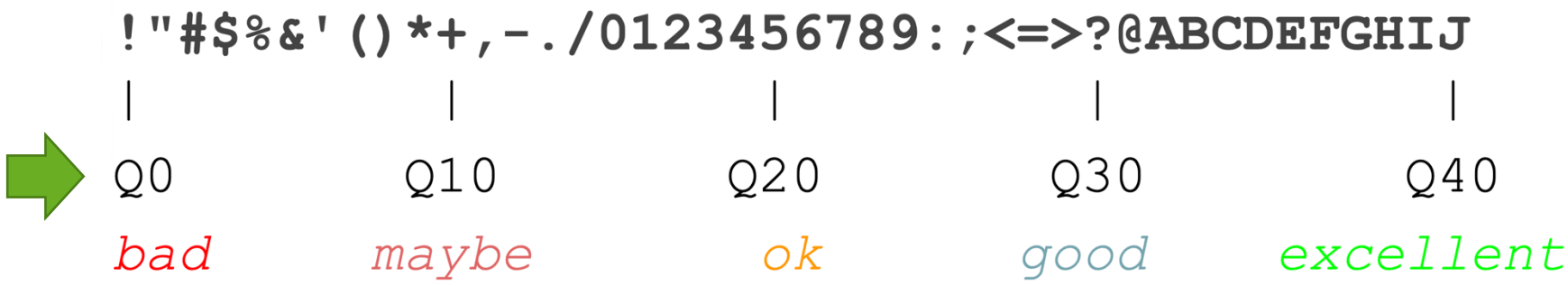
What is the meaning of these numbers?

Adapted from Joana Mourão (joam@dtu.dk); https://pt.wikipedia.org/wiki/Ficheiro:ASCII-Table.svg

Quality Score Encoding

Phred Quality Scores: Estimate of confidence in each base (a measure of reliability)

Phred Quality Score (Q)



Phred Quality (Q)	Chance it's wrong	Error probability (P)	Accuracy	Description
10	1 in 10	0.1	90%	**Maybe**
20	1 in 100	0.01	99%	**OK**
30	1 in 1000	0.001	99.9%	**Good**
40	1 in 10.000	0.0001	99.99%	**Very Good**
50	1 in 100.000	0.00001	99.999%	**Excellent**

A higher quality score is better
≥20 is considered **"acceptable"**

$$Q = -10 \log_{10} P \quad <=> \quad P = 10^{-Q/10}$$

Q = Phred quality score P = probability of base being incorrect

FastQC – Is my Sequence Data any Good?

Summary

- Basic Statistics
- Per base sequence quality
- Per sequence quality scores
- Per base sequence content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Adapter Content

Basic Statistics

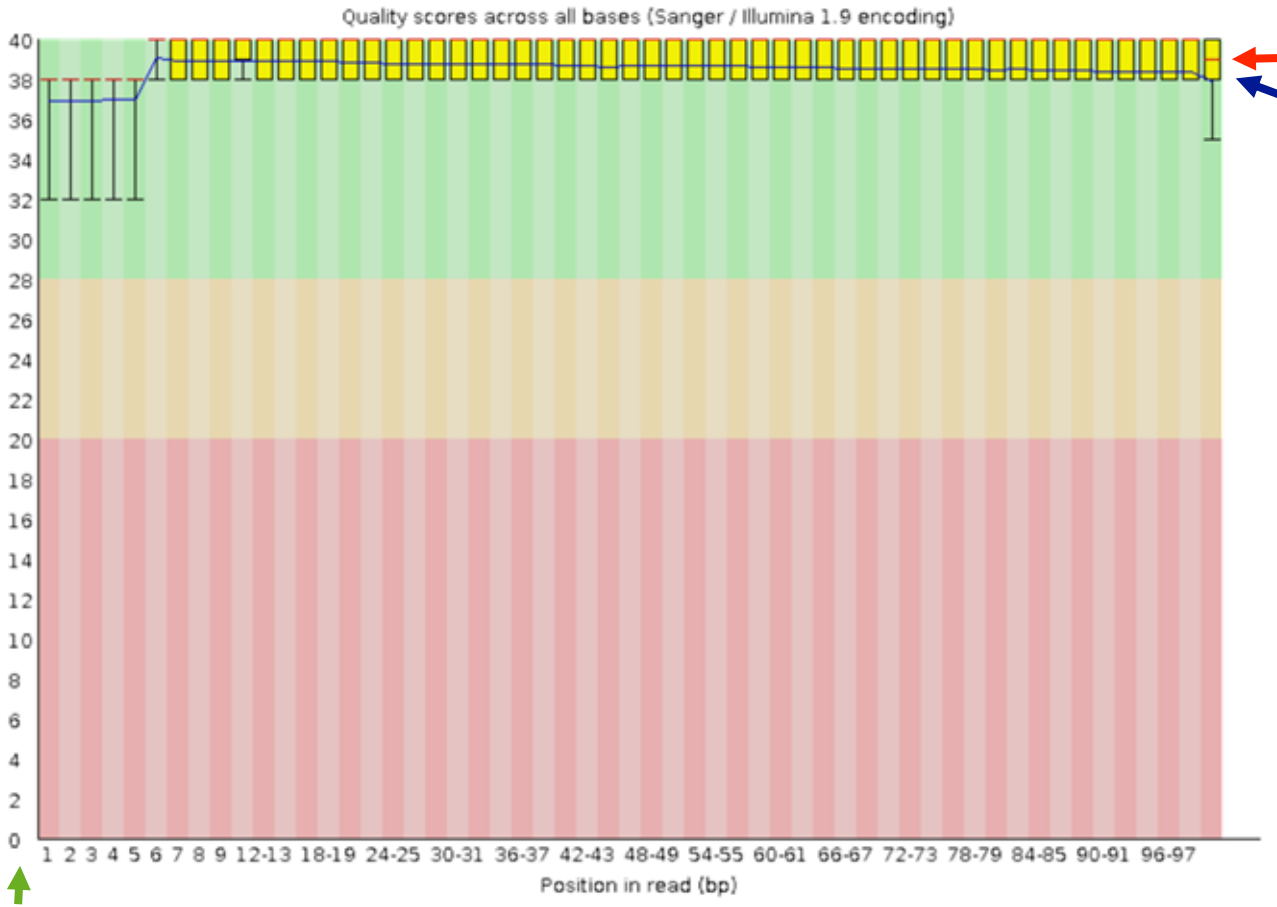
Measure	Value
Filename	SRR6052929_1.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	1502503
Sequences flagged as poor quality	0
Sequence length	35-101
%GC	50

FastQC – Is my Sequence Data any Good?

Summary

- Basic Statistics
- Per base sequence quality
- Per sequence quality scores
- Per base sequence content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Adapter Content

Per base sequence quality



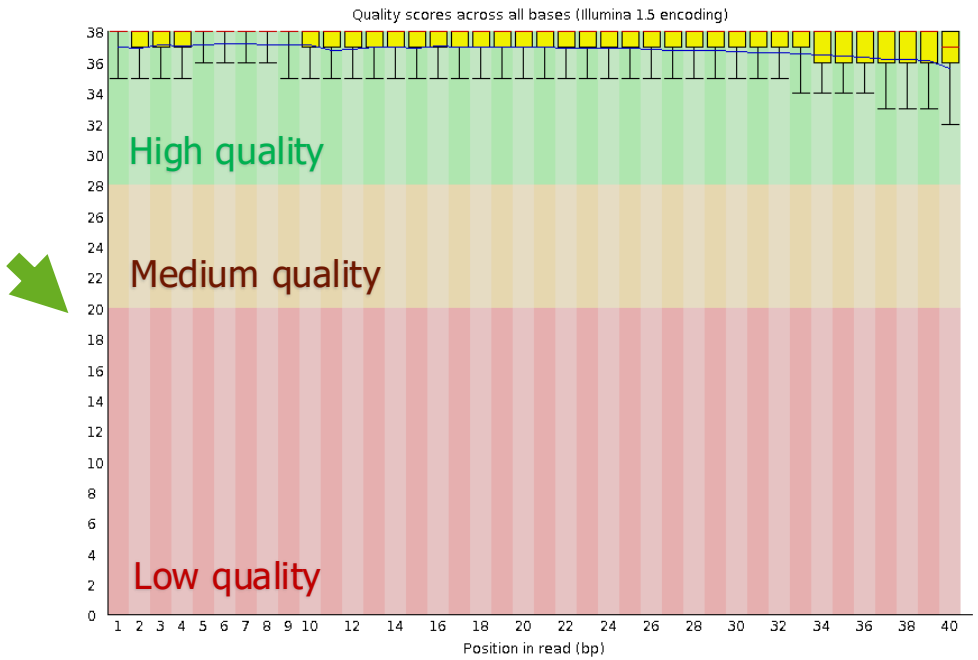
Median value (red)

Average quality (blue)

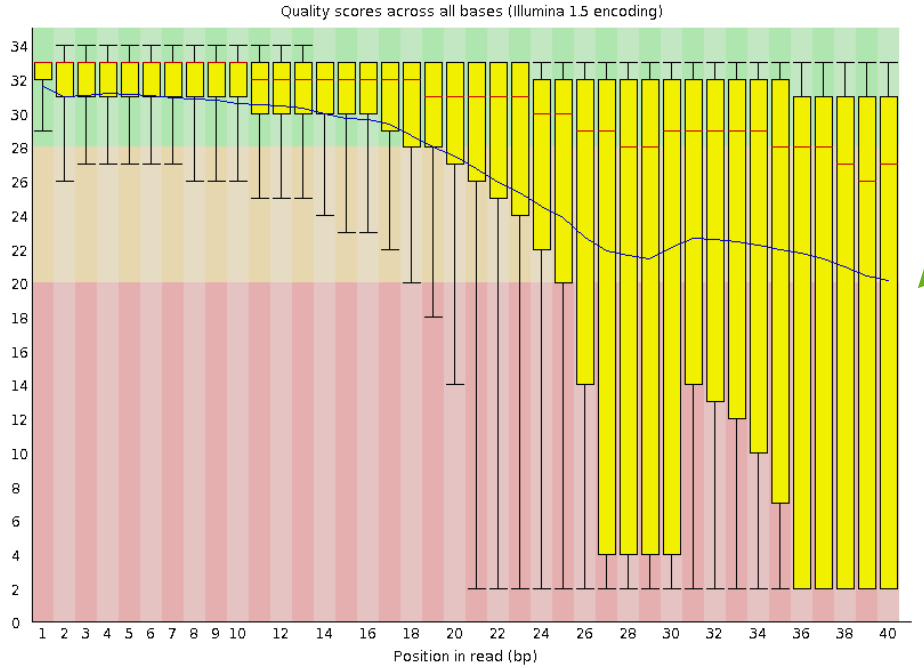
Y-axis is the "Phred" quality value (higher is better)

FastQC – Per Base Sequence Quality

A box-and-whisker plot showing aggregated **quality score** statistics at each position along all **reads** in the file (x-axis = read position, y-axis = Q-score)



Good Quality



Bad Quality

Possible approach (if needed): Trim from 3' to ≤ 20

Fastp Report

Summary

General

fastp version:	0.23.4 (https://github.com/OpenGene/fastp)
sequencing:	paired end (151 cycles + 151 cycles)
mean length before filtering:	146bp, 146bp
mean length after filtering:	146bp, 146bp
duplication rate:	1.051016%
Insert size peak:	230

Before filtering

total reads:	4.060834 M
total bases:	595.450771 M
Q20 bases:	539.283380 M (90.567249%)
Q30 bases:	486.205886 M (81.653414%)
GC content:	50.555634%

After filtering

total reads:	3.880680 M
total bases:	568.366141 M
Q20 bases:	522.457263 M (91.922658%)
Q30 bases:	472.771753 M (83.180844%)
GC content:	50.574606%

Filtering result

reads passed filters:	3.880680 M (95.563621%)
reads with low quality:	177.956000 K (4.382253%)
reads with too many N:	938 (0.023099%)
reads too short:	1.260000 K (0.031028%)

- Remove **adapters** & technical **sequences**
- Trim specific **low-quality bases** at 5′ (`--trim_front`) and 3′ (`--trim_tail`) ends
- **Filter** and **drop reads**: Ns, low-quality bases, or below minimum length
- **Correct overlapping** paired-end **reads**: in paired-end mode (`-c`), aligns and overlaps read pairs
- **Trim poly-G/X** tails

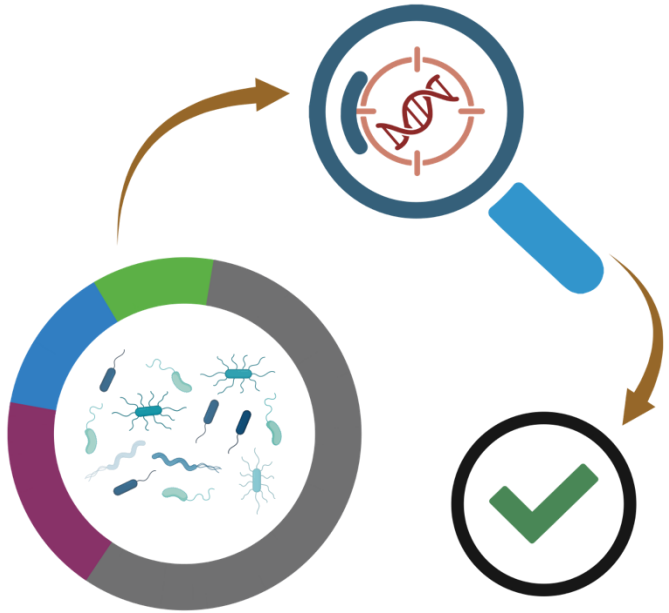
Assessment of Data Quality

SPECIES IDENTIFICATION

Command Line Tools - Contamination

Kraken/Bracken Inter-Species

Reads



GTDB-Tk Inter-Species

Assembly

Confindr (Intra-Species)

Reads

https://github.com/Ecogenomics/GTDBTk

https://github.com/DerrickWood/kraken2

https://github.com/OLC-Bioinformatics/ConFindr

KmerFinder – Species Identification

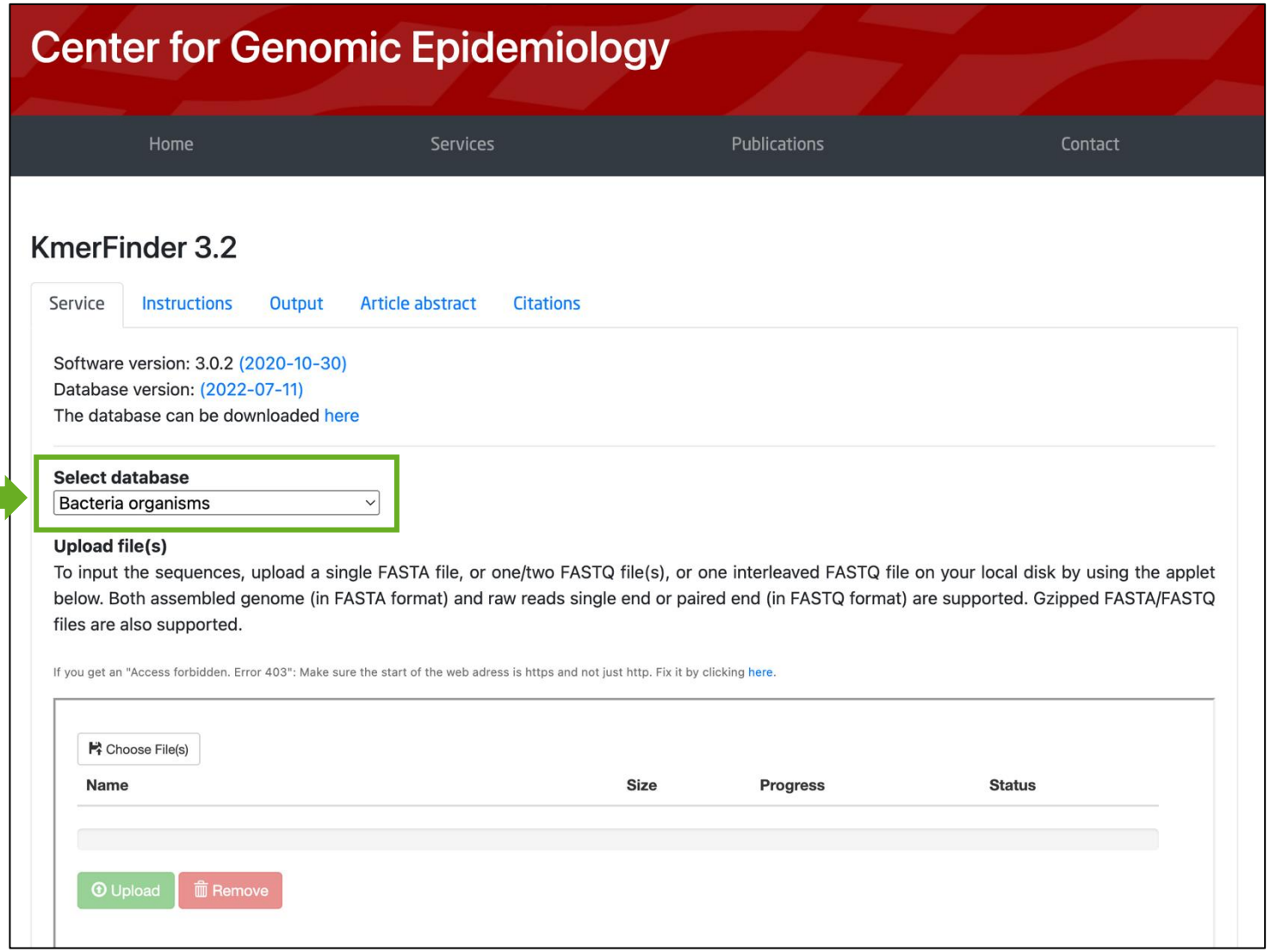
Fast species identification tool using k-mer matching algorithm

Uses 16-mers extracted from whole-genome sequencing data (FASTA/FASTQ format) compared against curated reference databases

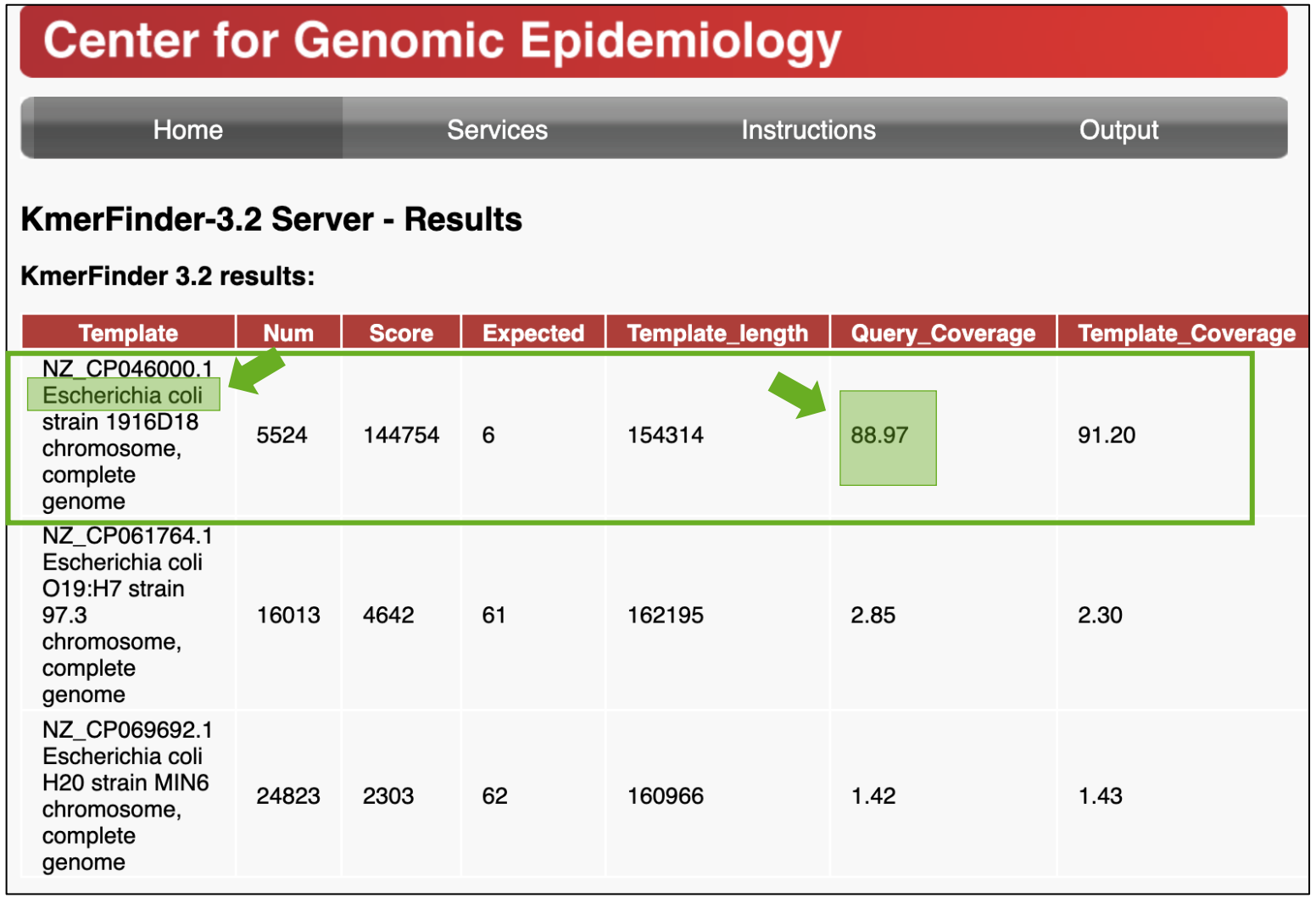
Free web service (also available as command-line tool)

https://cge.food.dtu.dk/services/KmerFinder/

KmerFinder - Upload



KmerFinder - Output



Center for Genomic Epidemiology

Home | Services | Instructions | Output

KmerFinder-3.2 Server - Results

KmerFinder 3.2 results:

Template	Num	Score	Expected	Template_length	Query_Coverage	Template_Coverage
NZ_CP046000.1 Escherichia coli strain 1916D18 chromosome, complete genome	5524	144754	6	154314	88.97	91.20
NZ_CP061764.1 Escherichia coli O19:H7 strain 97.3 chromosome, complete genome	16013	4642	61	162195	2.85	2.30
NZ_CP069692.1 Escherichia coli H20 strain MIN6 chromosome, complete genome	24823	2303	62	160966	1.42	1.43

Assessment of Sequencing Data Quality

Metrics and Available Tools

Quality Control Metrics (EURL-AMR)

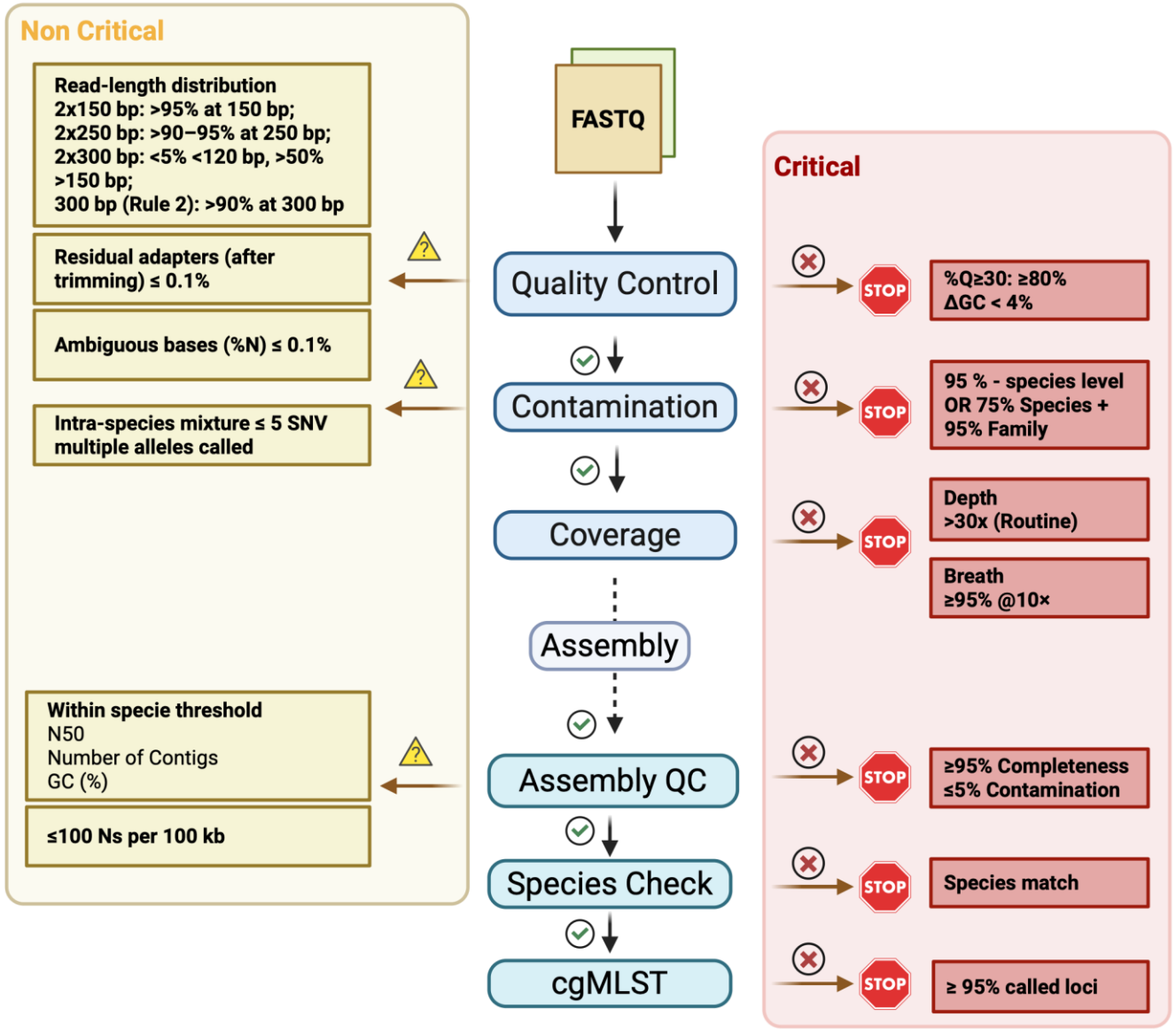
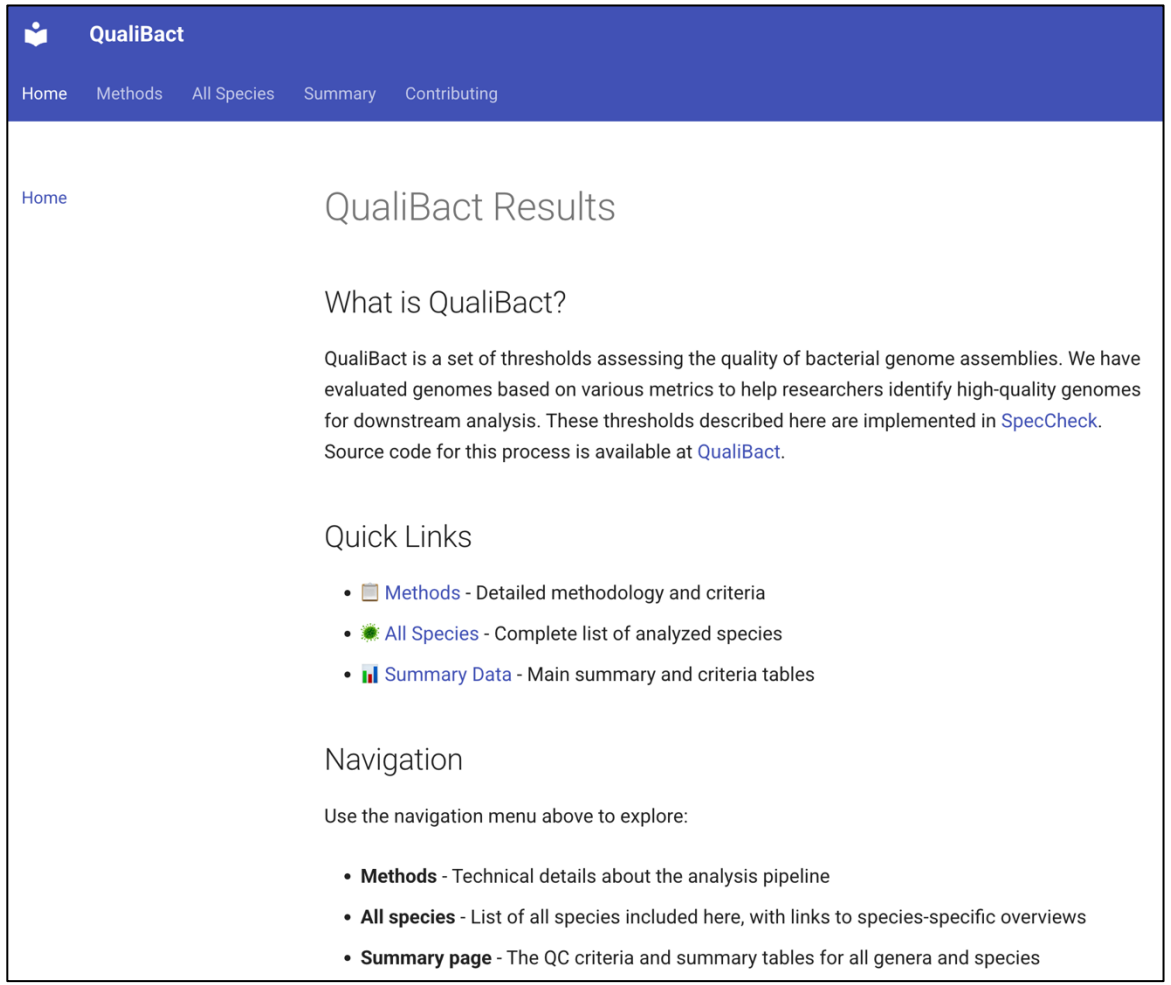
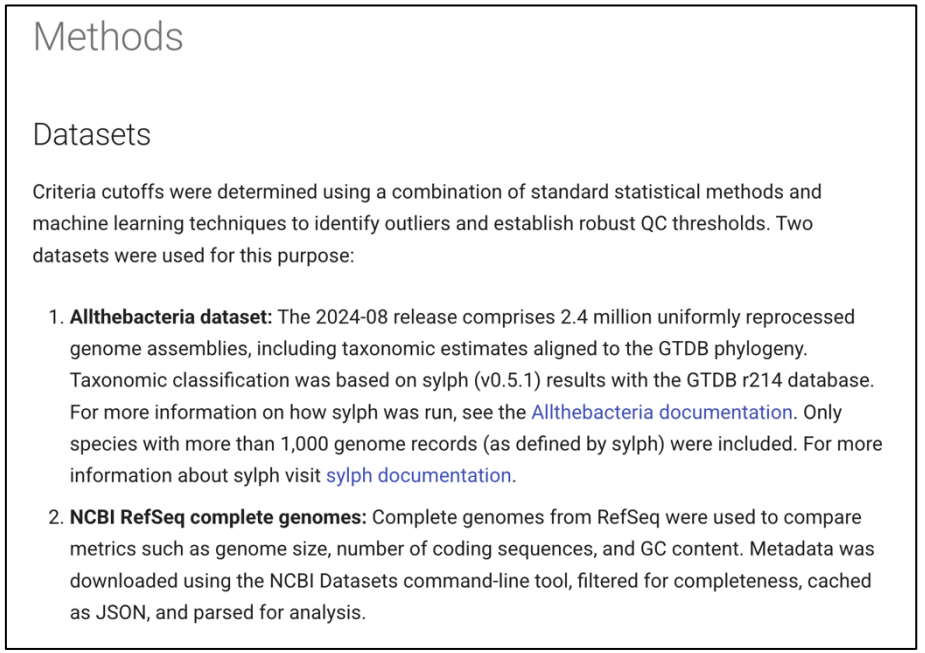


Image Created with BioRender.com

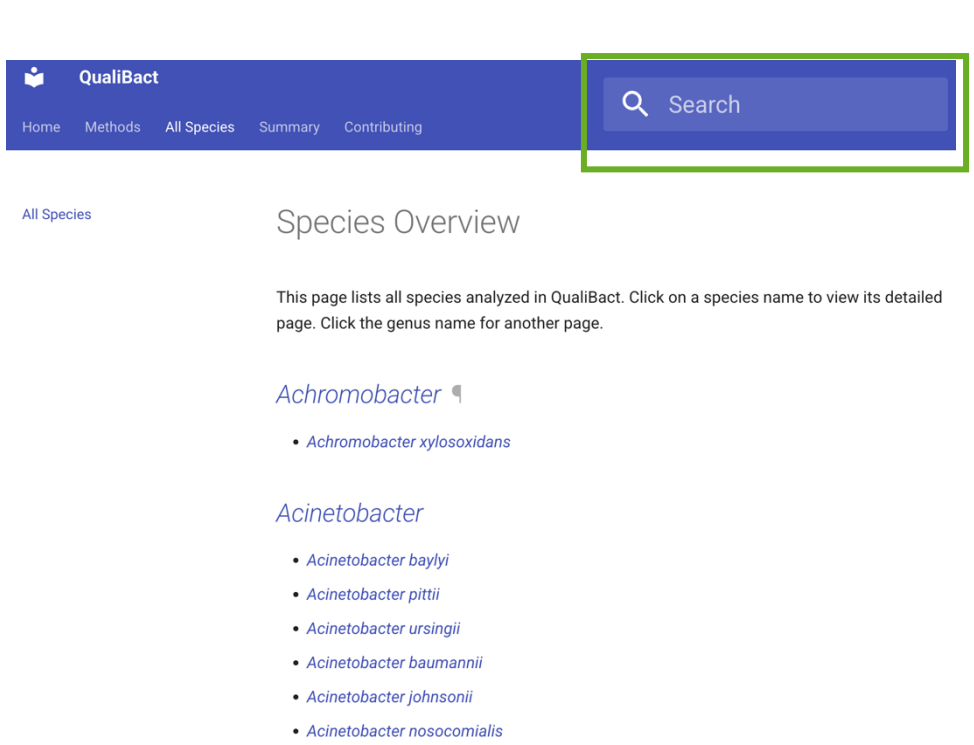
QualiBact - Species Specific Threshold





Adapted from Joana Mourão (joam@dtu.dk); https://happykhan.github.io/qualibact/

QualiBact - Species Specific Threshold



304 Species

Escherichia coli

This is the QualiBact page for *Escherichia coli*. For detailed methods on how these thresholds were calculated, please see Methods. The suggested thresholds are:

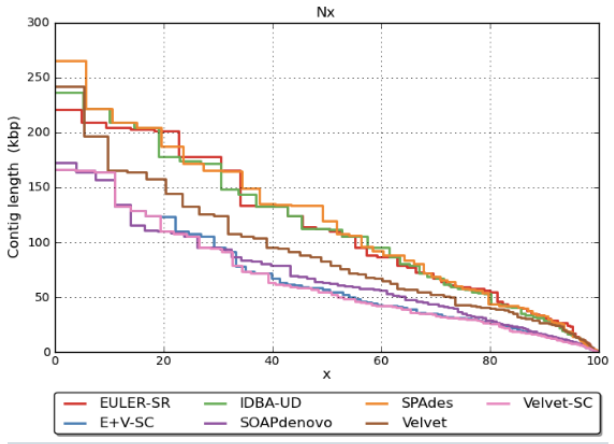
metric	lower_bounds	upper_bounds
N50	20000.0	
no_of_contigs		700.0
GC_Content	50.0	52.0
Completeness	95.0	
Contamination		16.0
Total_Coding_Sequences	3900.0	6500.0
Genome_Size	4100000.0	6300000.0

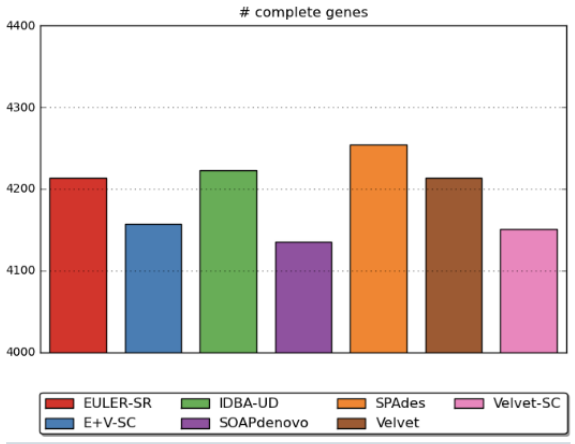
QualiBact

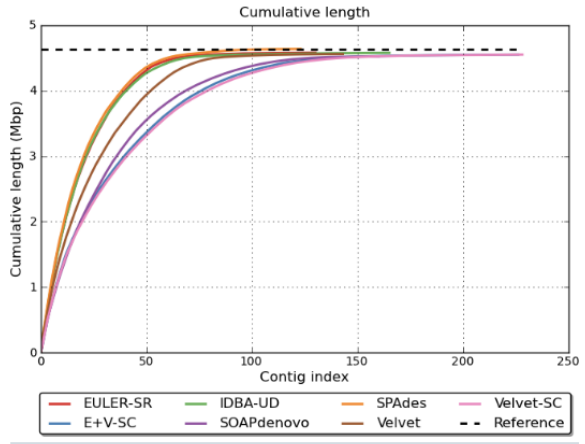
QUAST – Quality Assessment Tool

Provides comprehensive **metrics** and visual reports to assess the **accuracy**, **completeness**, and **contiguity** of assemblies → suitable for *de novo* **short-read**, **long-read**, and **hybrid** assemblies

Samples of QUAST plots:







https://quast.sourceforge.net/webquast.html

Gurevich et al. (2013). Bioinformatics. 29(8):1072–1075

Genome Fraction

All statistics are based on contigs of size >= 500 bp, unless otherwise noted (e.g., "# contigs (>= 0 bp)" and "Total length (>= 0 bp)" include all contigs).

Aligned to "
5372 genomic features

5 099 562 bp | 7 fragments | 50.55% G+C

Worst Median Best ☑ Show heatmap

Alignment-based statistics			
Genome fraction (%)	100	97.869	98.327
Duplication ratio	1	1	1.001
# genomic features	5370 + 2 part	5154 + 66 part	5214 + 74 part
Largest alignment	4 765 154	522 328	523 482
Total aligned length	5 100 438	4 992 703	5 018 679
NGA50	4 765 154	156 856	159 015
LGA50	1	11	11
Misassemblies			
# misassemblies	0	1	2
Misassembled contigs length	0	3281	7614
Per base quality			
# mismatches per 100 kbp	2.25	0.9	0.36
# indels per 100 kbp	0.22	0.1	0.08
# N's per 100 kbp	0	0	0
Statistics without reference			
# contigs	8	97	106
Largest contig	4 765 154	522 328	523 482
Total length	5 102 006	4 994 158	5 020 538
Total length (>= 1000 bp)	5 102 006	4 985 757	5 005 545
Total length (>= 10000 bp)	5 090 218	4 828 134	4 868 493
Total length (>= 50000 bp)	5 090 218	4 361 422	4 421 478

Extended report

Genome Fraction (%): percentage of the reference genome that is covered by the assembled contigs → related to the **Total Aligned Length**

Gurevich et al. (2013). Bioinformatics. 29(8):1072–1075

Genomic Features

All statistics are based on contigs of size >= 500 bp, unless otherwise noted (e.g., "# contigs (>= 0 bp)" and "Total length (>= 0 bp)" include all contigs).

Aligned to "
5372 genomic features

5 099 562 bp | 7 fragments | 50.55% G+C

Worst Median Best

Show heatmap

Alignment-based statistics			
Genome fraction (%)	100	97.869	98.327
Duplication ratio	1	1	1.001
# genomic features	5370 + 2 part	5154 + 66 part	5214 + 74 part
Largest alignment	4 765 154	522 328	523 482
Total aligned length	5 100 438	4 992 703	5 018 679
NGA50	4 765 154	156 856	159 015
LGA50	1	11	11
Misassemblies			
# misassemblies	0	1	2
Misassembled contigs length	0	3281	7614
Per base quality			
# mismatches per 100 kbp	2.25	0.9	0.36
# indels per 100 kbp	0.22	0.1	0.08
# N's per 100 kbp	0	0	0
Statistics without reference			
# contigs	8	97	106
Largest contig	4 765 154	522 328	523 482
Total length	5 102 006	4 994 158	5 020 538
Total length (>= 1000 bp)	5 102 006	4 985 757	5 005 545
Total length (>= 10000 bp)	5 090 218	4 828 134	4 868 493
Total length (>= 50000 bp)	5 090 218	4 361 422	4 421 478

Extended report

 # Genomic features: how many annotated elements from a reference (e.g., genes, coding sequences, rRNAs, tRNAs) are found in your assembly → either **fully** or **partially**

Gurevich et al. (2013). Bioinformatics. 29(8):1072–1075

Adapted from Joana Mourão (joam@dtu.dk); https://github.com/ablab/quast

Mismatches and Indels

All statistics are based on contigs of size >= 500 bp, unless otherwise noted (e.g., "# contigs (>= 0 bp)" and "Total length (>= 0 bp)" include all contigs).

Aligned to "
5372 genomic features

5 099 562 bp | 7 fragments | 50.55% G+C

Worst Median Best

Show heatmap

Alignment-based statistics			
Genome fraction (%)	100	97.869	98.327
Duplication ratio	1	1	1.001
# genomic features	5370 + 2 part	5154 + 66 part	5214 + 74 part
Largest alignment	4 765 154	522 328	523 482
Total aligned length	5 100 438	4 992 703	5 018 679
NGA50	4 765 154	156 856	159 015
LGA50	1	11	11
Misassemblies			
# misassemblies	0	1	2
Misassembled contigs length	0	3281	7614
Per base quality			
# mismatches per 100 kbp	2.25	0.9	0.36
# indels per 100 kbp	0.22	0.1	0.08
# N's per 100 kbp	0	0	0
Statistics without reference			
# contigs	8	97	106
Largest contig	4 765 154	522 328	523 482
Total length	5 102 006	4 994 158	5 020 538
Total length (>= 1000 bp)	5 102 006	4 985 757	5 005 545
Total length (>= 10000 bp)	5 090 218	4 828 134	4 868 493
Total length (>= 50000 bp)	5 090 218	4 361 422	4 421 478

Extended report

Mismatches and Indels: average number of base mismatches or insertions/deletions per 100.000 bp → **lower number** suggests **higher accuracy**

Gurevich et al. (2013). Bioinformatics. 29(8):1072–1075

Number of Contigs

All statistics are based on contigs of size >= 500 bp, unless otherwise noted (e.g., "# contigs (>= 0 bp)" and "Total length (>= 0 bp)" include all contigs).

Aligned to "
5372 genomic features

5 099 562 bp | 7 fragments | 50.55% G+C

Worst Median Best

Show heatmap

Alignment-based statistics			
Genome fraction (%)	100	97.869	98.327
Duplication ratio	1	1	1.001
# genomic features	5370 + 2 part	5154 + 66 part	5214 + 74 part
Largest alignment	4 765 154	522 328	523 482
Total aligned length	5 100 438	4 992 703	5 018 679
NGA50	4 765 154	156 856	159 015
LGA50	1	11	11
Misassemblies			
# misassemblies	0	1	2
Misassembled contigs length	0	3281	7614
Per base quality			
# mismatches per 100 kbp	2.25	0.9	0.36
# indels per 100 kbp	0.22	0.1	0.08
# N's per 100 kbp	0	0	0
Statistics without reference			
# contigs	8	97	106
Largest contig	4 765 154	522 328	523 482
Total length	5 102 006	4 994 158	5 020 538
Total length (>= 1000 bp)	5 102 006	4 985 757	5 005 545
Total length (>= 10000 bp)	5 090 218	4 828 134	4 868 493
Total length (>= 50000 bp)	5 090 218	4 361 422	4 421 478

Extended report

Short-read sequencing:

Difficult to obtain a complete and closed bacterial genome

#contigs (>= 500 bp): total number of contigs in the assembly → **fewer contigs** generally suggest **better contiguity** → gene identification and annotation are more feasible

Often, contigs below a certain threshold are not counted (e.g., 200 bp or 500 bp)

Gurevich et al. (2013). Bioinformatics. 29(8):1072–1075

Adapted from Joana Mourão (joam@dtu.dk); https://github.com/ablab/quast

Total Length

All statistics are based on contigs of size >= 500 bp, unless otherwise noted (e.g., "# contigs (>= 0 bp)" and "Total length (>= 0 bp)" include all contigs).

Aligned to "
5372 genomic features

5 099 562 bp | 7 fragments | 50.55% G+C

Worst Median Best

Show heatmap

Alignment-based statistics			
Genome fraction (%)	100	97.869	98.327
Duplication ratio	1	1	1.001
# genomic features	5370 + 2 part	5154 + 66 part	5214 + 74 part
Largest alignment	4 765 154	522 328	523 482
Total aligned length	5 100 438	4 992 703	5 018 679
NGA50	4 765 154	156 856	159 015
LGA50	1	11	11
Misassemblies			
# misassemblies	0	1	2
Misassembled contigs length	0	3281	7614
Per base quality			
# mismatches per 100 kbp	2.25	0.9	0.36
# indels per 100 kbp	0.22	0.1	0.08
# N's per 100 kbp	0	0	0
Statistics without reference			
# contigs	8	97	106
Largest contig	4 765 154	522 328	523 482
Total length	5 102 006	4 994 158	5 020 538
Total length (>= 1000 bp)	5 102 006	4 985 757	5 005 545
Total length (>= 10000 bp)	5 090 218	4 828 134	4 868 493
Total length (>= 50000 bp)	5 090 218	4 361 422	4 421 478

Extended report

- Total **length** of **all contigs** after the assembly
- For whole genome sequencing we expect it to be **close** to the **actual size** of the **genome**
- Much larger or smaller length than expected → may indicate possible **contamination** or **wrong species**

Rule of thumb:
Good assembly – usually within 5–10% of the species' known genome size

Gurevich et al. (2013). Bioinformatics. 29(8):1072–1075

GC Percentage

All statistics are based on contigs of size >= 500 bp, unless otherwise noted (e.g., "# contigs (>= 0 bp)" and "Total length (>= 0 bp)" include all contigs).

Aligned to 5 099 562 bp | 7 fragments | 50.55% G+C
5372 genomic features

Worst Median Best ☑ Show heatmap

Statistics without reference			
# contigs	8	97	106
# contigs (>= 0 bp)	13	102	181
# contigs (>= 1000 bp)	8	85	85
# contigs (>= 5000 bp)	6	59	57
# contigs (>= 10000 bp)	5	46	47
# contigs (>= 25000 bp)	5	37	37
# contigs (>= 50000 bp)	5	28	29
Largest contig	4 765 154	522 328	523 482
Total length	5 102 006	4 994 158	5 020 538
Total length (>= 0 bp)	5 103 319	4 996 528	5 040 462
Total length (>= 1000 bp)	5 102 006	4 985 757	5 005 545
Total length (>= 5000 bp)	5 096 296	4 917 836	4 945 228
Total length (>= 10000 bp)	5 090 218	4 828 134	4 868 493
Total length (>= 25000 bp)	5 090 218	4 688 275	4 705 006
Total length (>= 50000 bp)	5 090 218	4 361 422	4 421 478
N50	4 765 154	168 995	159 015
N90	4 765 154	43 746	42 118
auN	4 456 118	192 978	192 990
L50	1	10	11
L90	1	31	32
GC (%)	50.55	50.5	50.5

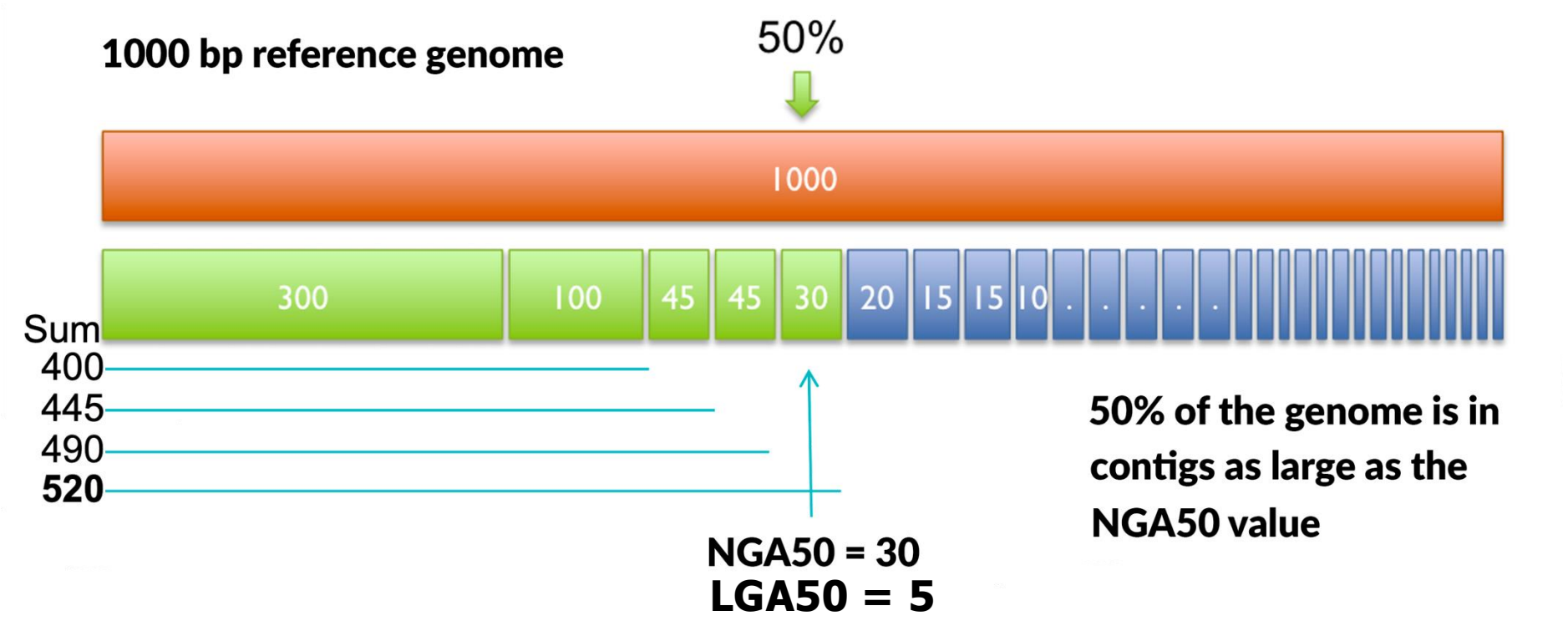
- The percentage of **guanine** (G) and **cytosine** (C) bases in the genome
- It is usually **stable within** a **species**
- **Atypical GC** content for the species → possible contamination or mislabeling

Gurevich et al. (2013). Bioinformatics. 29(8):1072–1075

NGA50 and LGA50 values

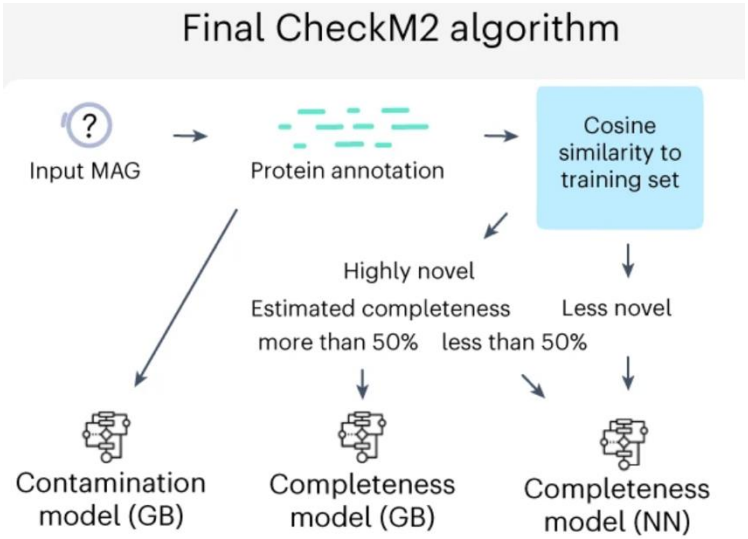
NGA50: the length of the contig at which 50% of the genome is covered by contigs of that length or larger → **higher NGA50, better assembly continuity**

LGA50: the smallest number of aligned contigs that together constitute 50% of the genome size → **smaller LGA50, better assembly continuity**



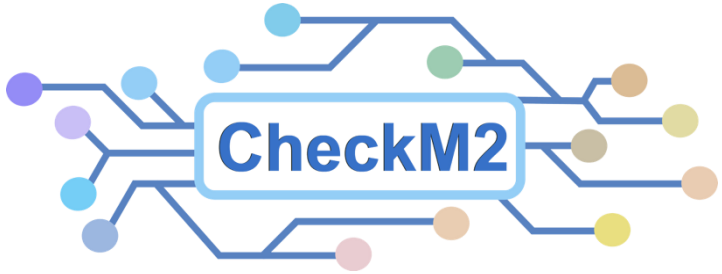
CheckM2

Designed to assess the quality (**completeness** and **contamination**) of genome assemblies, based on the presence of hundreds of genomic features - such as protein family counts, metabolic pathway/module completeness, amino-acid composition, single- and multi-copy genes - contrarily to other tools that only use single-copy marker genes (e.g., CheckM1 or BUSCO)



Parks et al. (2015). Genome Res. 25(7):1043-55; https://github.com/Ecogenomics/CheckM/; https://github.com/chklovski/CheckM2

Adapted from Joana Mourão (joam@dtu.dk); https://github.com/chklovski/CheckM2



A high-quality genome typically has **>90% completeness** and **<5% contamination**

Name	Completeness	Contamination	Completeness_Model_Used	Coding_Density	Contig_N50	Average_Gene_Length	Genome_Size	GC_Content	Total_Coding_Sequences	Total_Contigs	Max_Contig_Length
Strain-01	100	0.3	Neural Network (Specific Model)	0.876	5229427	311.8831068	5494292	0.5	5150	6	5229427
Strain-02	100	0.26	Neural Network (Specific Model)	0.879	5159857	311.3406763	5490316	0.5	5175	9	5159857
Strain-03	100	0.22	Neural Network (Specific Model)	0.835	2783198	307.9881094	2786908	0.33	2523	2	2783198
Strain-04	100	1.38	Neural Network (Specific Model)	0.834	2819759	300.8207441	2873979	0.33	2661	2	2819759
Strain-05	100	0.09	Neural Network (Specific Model)	0.878	2825666	313.6127737	2932174	0.37	2740	3	2825666
Strain-06	100	1.49	Neural Network (Specific Model)	0.847	2777239	299.7961062	2995648	0.38	2825	3	2777239

Completeness (%): The estimated fraction of the "expected" gene/content repertoire that is present in your assembly; a value near 100 % means nearly all "expected" features were recovered.

Contamination (%): The estimated amount of "extra" (likely foreign or redundant) content, based on over-represented gene families or conflicting pathway signals.

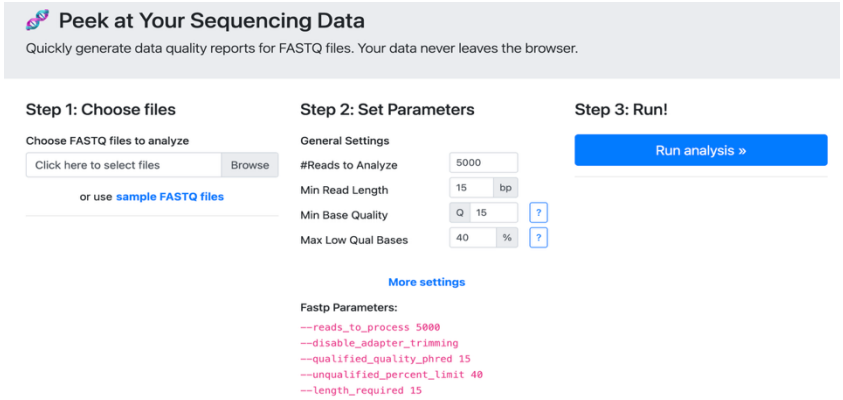
Parks et al. (2015). Genome Res. 25(7):1043-55; https://github.com/Ecogenomics/CheckM/; https://github.com/chklovski/CheckM2

Adapted from Joana Mourão (joam@dtu.dk); https://github.com/chklovski/CheckM2

Web-based tools

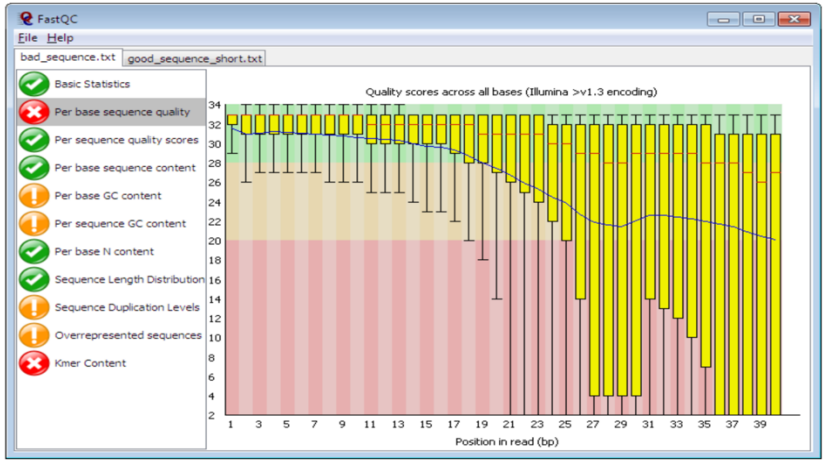
Raw Reads Quality Control

https://fastq.bio/



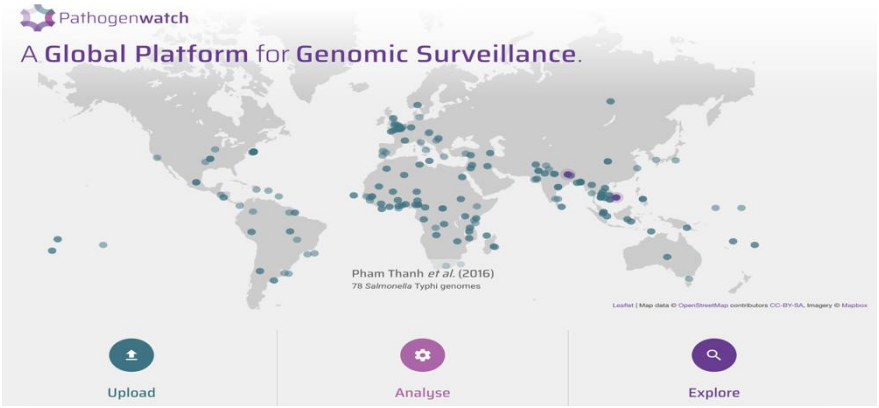
Raw Reads Quality Control

https://www.bioinformatics.babraham.ac.uk/projects/fastqc/



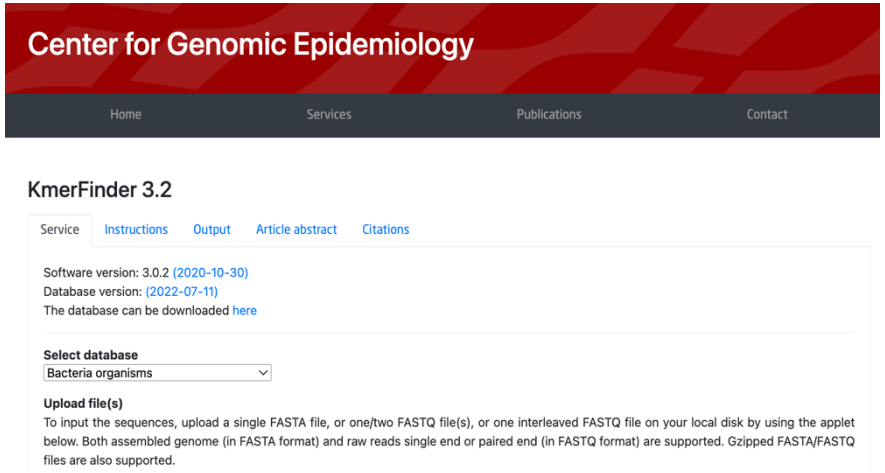
Assembly QC cgMLST

https://pathogen.watch/

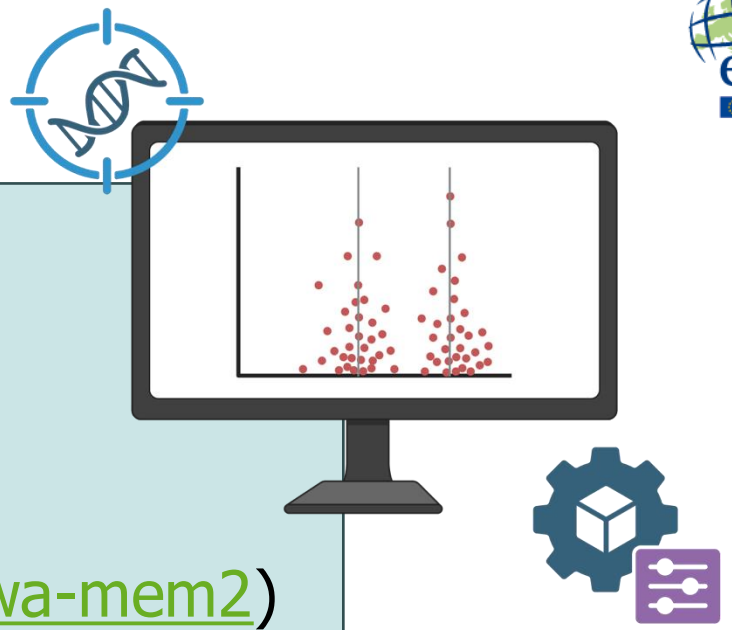


Contamination

https://cge.food.dtu.dk/services/KmerFinder-3.2/



Command Line tools



- FastP (https://github.com/OpenGene/fastp)
- FastQC (https://github.com/OpenGene/fastp)
- Samtools (https://www.htslib.org/)
- bwa-mem2 (https://github.com/bwa-mem2/bwa-mem2)
- Kraken2 (https://github.com/DerrickWood/kraken2)
- Bracken (https://github.com/jenniferlu717/Bracken)
- CheckM2 (https://github.com/chklovski/CheckM2)
- GTDB-Tk (https://github.com/Ecogenomics/GTDBTk)
- chewBBACA (https://github.com/B-UMMI/chewBBACA)

Questions

Acknowledgements

The creation of this training material was commissioned by ECDC to Technical University of Denmark (DTU) with the direct involvement of João Cardoso (Research Assistant at DTU, Msc – joacar@food.dtu.dk)