

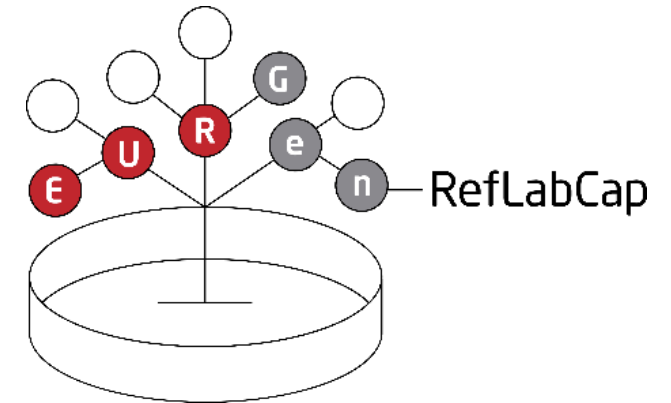
# EURGen-RefLabCap

## Technical training workshop # 1

Second day

Wednesday, 7 December 2022

09:00 – 17:00 CET



# Previously...

## **First day (virtual) – Tuesday 29 November 2022, 10:00 - 12:30 CET**

10:00 - 10:15: Introduction and agenda for the day (Ana Rita Rebelo, DTU)

10:15 - 11:00: From isolate to WGS - biochemical principles (Ana Rita Rebelo, DTU)

11:00 - 11:15: Coffee break

11:15 - 12:00: From isolate to WGS - WGS data and bioinformatics (Jette Sejer Kjeldgaard, DTU)

12:00 - 12:30: Quality control of WGS data (Ana Rita Rebelo, DTU)

## Second day (physical) – Wednesday 7 December 2022, 9:00 - 17:00 CET

9:00 - 9:30: Introduction and agenda for the day (Ana Rita Rebelo, DTU)

9:30 - 15:00: Laboratory work - Illumina MiSeq library preparation and sequencing (including *ad hoc* coffee and lunch)

15:00 - 15:30: Coffee break

15:30 - 16:15: Exercise about quality control of WGS data (Ana Rita Rebelo, DTU)

16:15 - 17:00: Exercise about bioinformatics tools for species identification and subtyping (Jette Sejer Kjeldgaard, DTU)

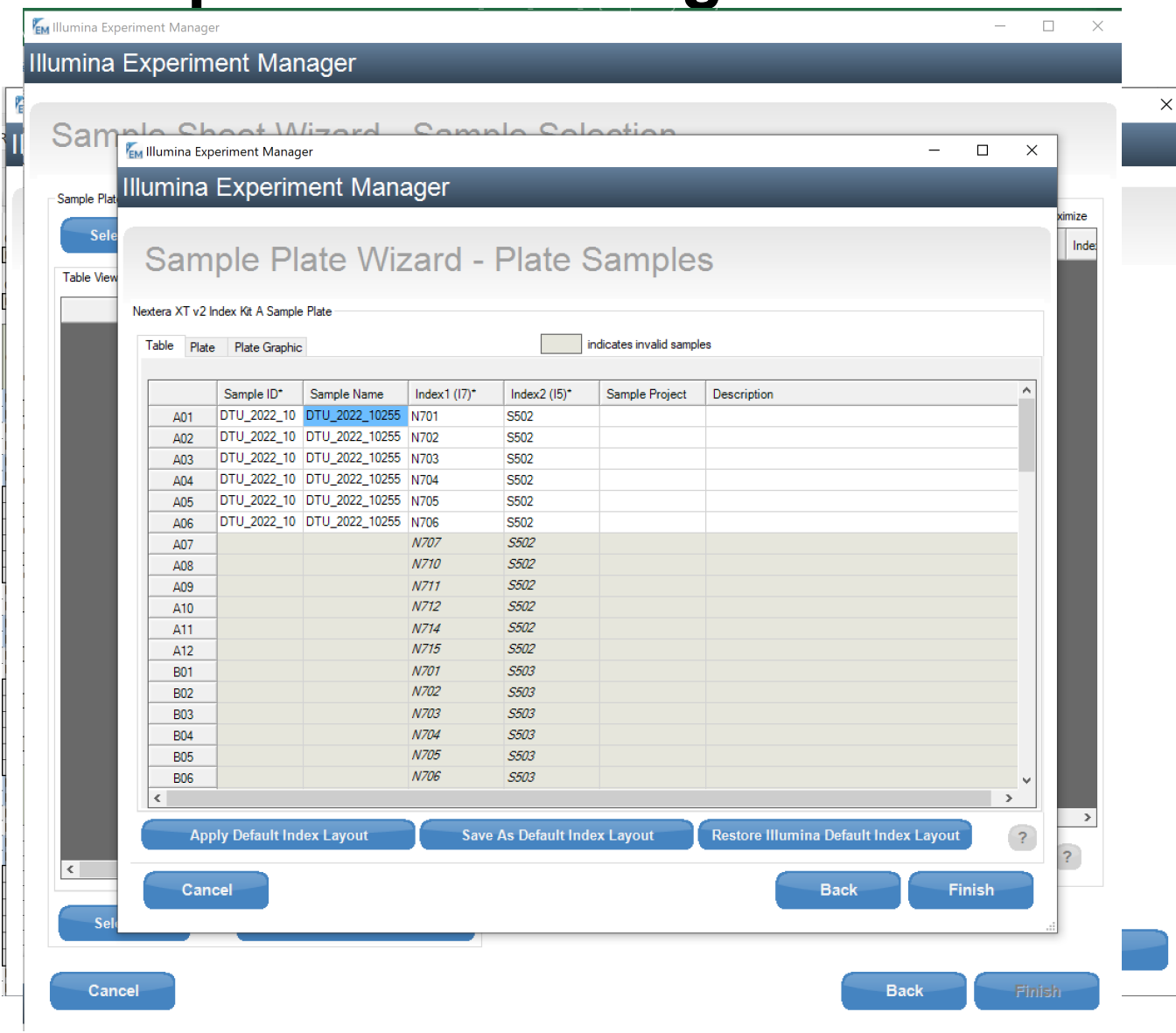
*[17:00: Bus transportation provided from DTU to the restaurant]*

17:45 : Project dinner at Restaurant Bistro Royal København (Kongens Nytorv 26, 1050 København K)

Laboratory work

# **Illumina MiSeq library preparation and sequencing**

# Illumina Experiment Manager



Sample Plate Wizard - Plate Samples

Nextera XT v2 Index Kit A Sample Plate

Table Plate Plate Graphic indicates invalid samples

	Sample ID*	Sample Name	Index1 (I7)*	Index2 (I5)*	Sample Project	Description
A01	DTU_2022_10	DTU_2022_10255	N701	S502		
A02	DTU_2022_10	DTU_2022_10255	N702	S502		
A03	DTU_2022_10	DTU_2022_10255	N703	S502		
A04	DTU_2022_10	DTU_2022_10255	N704	S502		
A05	DTU_2022_10	DTU_2022_10255	N705	S502		
A06	DTU_2022_10	DTU_2022_10255	N706	S502		
A07			N707	S502		
A08			N710	S502		
A09			N711	S502		
A10			N712	S502		
A11			N714	S502		
A12			N715	S502		
B01			N701	S503		
B02			N702	S503		
B03			N703	S503		
B04			N704	S503		
B05			N705	S503		
B06			N706	S503		

Apply Default Index Layout Save As Default Index Layout Restore Illumina Default Index Layout ?

Cancel Back Finish

***Ad hoc* coffee and lunch available from 11:00**



# Coffee break

**Back at 15:30.**



# Pending questions from the first day

*“Would you comment on the EDTA concentration in Elution buffers in the context of common NGS library prep's first step - enzymatic fragmentation?”*

- Lopata A, Jójárt B, Surányi ÉV, Takács E, Bezúr L, Leveles I, Bendes ÁÁ, Viskolcz B, Vértessy BG, Tóth J. **Beyond Chelation: EDTA Tightly Binds Taq DNA Polymerase, MutT and dUTPase and Directly Inhibits dNTPase Activity**. Biomolecules. 2019 Oct 17;9(10):621. doi: 10.3390/biom9100621. PMID: 31627475; PMCID: PMC6843921.

**Other questions?**



Ana Rita Rebelo  
*anrire@food.dtu.dk*

# Exercise about quality control of WGS data – recap of first day

# From the first day of the workshop...

## About the library preparation procedures:

***Do not re-use materials – especially if you are not confident while doing it***

- Can you use the same tips to distribute the indexes during amplification?
- Can you use the same tips to distribute ethanol during library clean up?

***Pay attention to storage and thawing conditions***

- Why?

# From the first day of the workshop...

## Examples of nonsense results → Why are these specific cases nonsense?

- Species identification tool detects a high proportion of hits belonging to different species
- Different *bla* genes in the same position of the genome with <100% coverage
- Lack of known AMR determinants when there is phenotypic resistance
- Zero SNPs between isolates very separated according to metadata (time/space)

# From the first day of the workshop...

- Species identification tool detects a high proportion of hits belonging to different species

#Template	Num	Score	Expected	Template _length	Query_Covera ge	Template_Cov erage	Depth	tot_query_Covera ge	tot_template _Coverage	tot_dept h	q_value	p_value
NZ_CP020052.1 Proteus mirabilis strain AR_0059, complete genom	2942	125008	32	153944	50.44	81.98	0.81	50.44	81.98	0.81	124911.52	1.0e-26
NZ_CP028956.1 Morganella morganii strain AR_0133 chromosome	2930	26884	67	135302	10.85	20.30	0.20	11.50	21.20	0.21	26683.09	1.0e-26
NZ_CP029133.1 Proteus mirabilis strain AR379 chromosome, comp	2950	7336	86	154859	2.96	4.82	0.05	49.05	79.18	0.78	7081.89	1.0e-26
NZ_CP023505.1 Morganella morganii strain FDAARGOS_365 chrom	2927	5672	73	129827	2.29	4.44	0.04	11.19	21.35	0.21	5456.59	1.0e-26
NZ_CP015347.1 Proteus mirabilis strain AOUC-001, complete geno	2941	2982	87	156373	1.20	1.92	0.02	50.07	81.05	0.79	2728.91	1.0e-26
NZ_CP014026.2 Morganella morganii strain FDAARGOS_172 chrom	2926	1264	71	124799	0.51	1.03	0.01	11.15	22.17	0.22	1064.13	1.0e-26

**Contamination!**

# From the first day of the workshop...

- Different *bla* genes in the same position of the genome with <100% coverage

escherichia coli complete			
Antimicrobial	Class	WGS-predicted phenotype	Genetic background
amikacin	aminoglycoside	No resistance	
tobramycin	aminoglycoside	No resistance	
gentamicin	aminoglycoside	No resistance	
cefepime	beta-lactam	Resistant	blaOXA-162 (blaOXA-162_GU197550), blaTEM-29 (blaTEM-29_DQ269440)
piperacillin+tazobactam	beta-lactam	Resistant	blaOXA-162 (blaOXA-162_GU197550), blaTEM-122 (blaTEM-122_AY307100)
cefoxitin	beta-lactam	No resistance	
ampicillin	beta-lactam	Resistant	blaTEM-1B (blaTEM-1B_AY458016), blaOXA-162 (blaOXA-162_GU197550), blaTEM-29 (blaTEM-29_DQ269440), blaTEM-122 (blaTEM-122_AY307100), blaTEM-55 (blaTEM-55_DQ286729), blaTEM-141 (blaTEM-141_AY956335), blaTEM-57 (blaTEM-57_FJ405211), blaTEM-1C (blaTEM-1C_FJ560503), blaTEM-135 (blaTEM-135_GQ896333)
ampicillin+clavulanic acid	beta-lactam	Resistant	blaTEM-122 (blaTEM-122_AY307100)
cefotaxime	beta-lactam	Resistant	blaTEM-29 (blaTEM-29_DQ269440)
imipenem	beta-lactam	Resistant	blaOXA-162 (blaOXA-162_GU197550)
ertapenem	beta-lactam	Resistant	blaOXA-162 (blaOXA-162_GU197550)
ceftazidime	beta-lactam	Resistant	blaTEM-29 (blaTEM-29_DQ269440)
temocillin	beta-lactam	No resistance	
meropenem	beta-lactam	Resistant	blaOXA-162 (blaOXA-162_GU197550)

# From the first day of the workshop...

- Different *bla* genes in the same position of the genome with <100% coverage

blaTEM-122	99.8838559814	861/861	1..861	NODE_96_lengt h_8473_cov_26. 387970	7452..8312	amoxicillin, amoxi cillin+clavulanic acid, ampicillin, a mpicillin+clavulan ic acid, piperacillin, p iperacillin+tazoba ctam, ticarcillin, tic arcillin+clavulanic acid
blaTEM-55	99.8838559814	861/861	1..861	NODE_96_lengt h_8473_cov_26. 387970	7452..8312	amoxicillin, ampici llin, cephalothin, pi peracillin, ticarcilli n
blaTEM-209	99.8838559814	861/861	1..861	NODE_96_lengt h_8473_cov_26. 387970	7452..8312	unknown beta-lac tams
blaTEM-1B	99.8838559814	861/861	1..861	NODE_96_lengt h_8473_cov_26. 387970	7452..8312	amoxicillin, ampici llin, cephalothin, pi peracillin, ticarcilli n
blaTEM-141	99.8838559814	861/861	1..861	NODE_96_lengt h_8473_cov_26. 387970	7452..8312	amoxicillin, ampici llin, cephalothin, pi peracillin, ticarcilli n
blaOXA-162	100.0	798/798	1..798	NODE_163_lengt h_2231_cov_6.1 64122	26..823	amoxicillin, ampici llin, cefepime, erta penam, imipenem , meropenem, piper acillin, piperacilli n+tazobactam
blaTEM-57	99.8838559814	861/861	1..861	NODE_96_lengt h_8473_cov_26. 387970	7452..8312	amoxicillin, ampici llin, cephalothin, pi peracillin, ticarcilli n
blaTEM-29	99.8838559814	861/861	1..861	NODE_96_lengt h_8473_cov_26. 387970	7452..8312	amoxicillin, ampici llin, aztreonam, ce fepime, cefotaxim e, ceftazidime, ceft riaxone, piperacilli n, ticarcillin

Artifacts of the tool:  
There's no perfect match  
to the gene

OK!

# From the first day of the workshop...

- Lack of known AMR determinants when there is phenotypic resistance

	Colistin MIC (mg/L)	Colistin ARG/PM
<i>Enterobacter aerogenes</i>	0,5	None
<i>Enterobacter aerogenes</i>	0,5	None
<i>Enterobacter cloacae</i>	0,5	None
<i>Enterobacter asburiae</i>	16	None
<i>Enterobacter cloacae</i>	0,5	None
<i>Enterobacter cloacae</i>	64	None
<i>Enterobacter cloacae</i>	0,5	None
<i>Enterobacter cloacae</i>	2	None
<i>Enterobacter cloacae</i>	0,5	None
<i>Enterobacter cloacae</i>	8	None

Phenotypic resistance

## Database is incomplete

- It's very important to be familiar with the AMR determinants actually present in the databases
- Resistance and susceptibility should never be predicted from genotypic analysis
- Results should be reported as presence or absence of genes or point mutations

# From the first day of the workshop...

- Zero SNPs between isolates very separated according to metadata (time/space)

Lenski, R. Experimental evolution and the dynamics of adaptation and genome evolution in microbial populations. *ISME J* 11, 2181–2194 (2017). <https://doi.org/10.1038/ismej.2017.69>

## “Structure of the experiment and motivating questions

The long-term evolution experiment, or LTEE, is simple both conceptually and practically. Twelve populations were started the same ancestral strain of *Escherichia coli* in 1988. The ancestral strain has no plasmids or functional prophages, and *E. coli* is not naturally transformable, so there is no horizontal gene transfer. However, each population has millions of cells that provide a continual supply of new mutations. The populations are propagated in a glucose-limited minimal salts medium at 37 ° C by transferring 1% of the volume into fresh medium every day. The 100-fold dilution and resulting regrowth allows  $\log_2 100 \approx 6.7$  generations each day. Samples from each population are periodically stored at -80 ° C, where they are available for later study. Importantly, the frozen cells remain viable, such that changes in performance can be analyzed at later times; and when accidents occur, the populations are restarted from the most recent whole-population samples. As of this writing, the LTEE has passed 66 000 generations.”



## From the first day of the workshop...

- Zero SNPs between isolates very separated according to metadata (time/space)

And how many mutations have accumulated in the evolving genomes? [Tenaillon et al. \(2016\)](#) sequenced 264 clonal genomes from the LTEE (2 clones from 11 time points and 12 populations) and compared them with the ancestral genome to identify the mutations. In total, they found over 17 000 mutations, although most were in the six populations that evolved hypermutability. Clones sampled at generation 50 000 from populations that were not hypermutable averaged ~75 mutations in their genomes in contrast to the billion-plus mutation events that occurred in each population. Those 75 or so mutations also pale in comparison with the differences between *E. coli* and *Salmonella* ([Ochman et al., 1999](#)), or between two *E. coli* isolates from nature ([Dixit et al., 2015](#)). However, it is not surprising because the time scale of the LTEE, while long for an experiment, is a drop in the bucket of evolutionary time.

~20 years

~3.7 mutations/year

Ana Rita Rebelo  
*anrire@food.dtu.dk*

# Exercise about quality control of WGS data

# Recap of QC thresholds

Raw data QC	Assembled data QC
<b>Number of reads</b> Should be as high as possible. No assessed cut-off exist, but enough to obtain the desired coverage of the organism genome	<b>Size of assembled genome</b> <i>Enterobacterales</i> : 4.5 Mb - 5.5 Mb Deviation should not be higher than 10%
<b>Average read length</b> Should correspond to that expected from the sequencing platform and kit.  Illumina MiSeq avg read length = 300 bps Illumina NextSeq avg read length = 150 bps	<b>Total number of contigs</b> Should be less than 500
<b>Coverage</b> Should as a minimum be 30x, and preferably even higher	<b>N50</b> Should be over 15.000 bp

$$\text{Coverage} = \text{Number of reads} \times \frac{\text{Read length}}{\text{Genome size}}$$

Genera or group	Expected genome size (million bps)
<i>Enterobacterales</i> (excluding <i>Proteus</i> )	4.5 – 5.5
<i>Proteus</i>	4
<i>Pseudomonas</i>	6.5 – 7
<i>Staphylococcus</i>	2.5 – 2.8
<i>Streptococcus</i>	1.7 – 2.2

# Explanation

- 1) All clinically relevant bacterial isolates were collected from all hospitals in a region in one day
- 2) All isolates were sequenced using Illumina NextSeq 500
- 3) Bioinformatics analysis was performed with internal pipeline which includes FastQC

# Exercise

- Evaluate the QC parameters for the isolates and identify the problems.
- Decide what the next step would be:
  - Re-sequence the extracted DNA?
  - Re-extract the DNA from the same culture?
  - Prepare a subculture and repeat the whole process?

# Exercise

Isolate ID	Expected species	Output from Quality Control pipeline					
		Bases (MB)	Reads	N50	Number of contigs	Longest	Total BPs
Isolate A	Escherichia coli	994	8019656	288501	46	545016	4976587
Isolate B	Escherichia coli	405	2746488	254172	101	399093	5302962
Isolate C	Escherichia coli	385	2595090	173253	196	386355	10511210
Isolate D	Escherichia coli	1532	11072180	354248	57	629640	5060120
Isolate E	Escherichia coli	816	5875512	188239	66	556756	4779513
Isolate F	Escherichia coli	602	4292106	66436	1897	492856	7310556
Isolate G	Klebsiella pneumoniae	40	289816	2920	2346	29790	4994958
Isolate H	Yersinia enterocolitica	852	6069858	121701	73	245199	4636672
Isolate I	Staphylococcus aureus	134	921174	116189	51	232965	2809977
Isolate J	Staphylococcus aureus	151	1053706	135887	117	342404	5038817
Isolate K	Staphylococcus aureus	560	3935968	125616	43	504236	2746951
Isolate L	Staphylococcus aureus	446	3124582	539934	19	992722	2709295
Isolate M	Staphylococcus aureus	14	101884	1102	1501	20858	1545320
Isolate N	Streptococcus group C (Hemolytic)	185	1327182	73298	139	287028	4190334
Isolate O	Streptococcus group C (Hemolytic)	311	2292512	73298	58	182696	2093457
Isolate P	Enterococcus sp.	184	1331140	100659	220	449018	6889533

# Discussion

Isolate ID	Expected species	Output from Quality Control pipeline					
		Bases (MB)	Reads	N50	Number of contigs	Longest	Total BPs
Isolate A	Escherichia coli	994	8019656	288501	46	545016	4976587
Isolate B	Escherichia coli	405	2746488	254172	101	399093	5302962
Isolate C	Escherichia coli	385	2595090	173253	196	386355	10511210
Isolate D	Escherichia coli	1532	11072180	354248	57	629640	5060120
Isolate E	Escherichia coli	816	5875512	188239	66	556756	4779513
Isolate F	Escherichia coli	602	4292106	66436	1897	492856	7310556
Isolate G	Klebsiella pneumoniae	40	289815	2920	2346	29790	4994958
Isolate H	Yersinia enterocolitica	852	6069858	121701	73	245199	4636672
Isolate I	Staphylococcus aureus	134	921174	116189	51	232965	2809977
Isolate J	Staphylococcus aureus	151	1053706	135887	117	342404	5038817
Isolate K	Staphylococcus aureus	560	3935968	125616	43	504236	2746951
Isolate L	Staphylococcus aureus	446	3124582	539934	19	992722	2709295
Isolate M	Staphylococcus aureus	14	101884	1102	1501	20858	1545320
Isolate N	Streptococcus group C (Hemolytic)	185	1327182	73298	139	287028	4190334
Isolate O	Streptococcus group C (Hemolytic)	311	2292512	73298	58	182696	2093457
Isolate P	Enterococcus sp.	184	1331140	100659	220	449018	6889533

Isolate ID	Expected species	Output from Quality Control pipeline						Downstream analysis		
		Bases (MB)	Reads	N50	Nr. of contigs	Longest	Total BPs	Species	MLST	Resistance
<b>Isolate C</b>	Escherichia coli	385	2595090	173253	196	386355	10511210	Escherichia coli	ecoli[Unknown ST]	blaSHV-33,fosA-like,oqxA-like,oqxB-like
<b>Isolate C - reextract</b>	Escherichia coli	149	1082756	171241	90	452935	5117962	Escherichia coli	ecoli[ST-73]	

- There was a contamination – likely with another *E. coli* isolate
- In case of contaminations, DNA should always be re-extracted (and not re-run)
- After re-extracting new DNA and sequencing, there is good quality of results but loss of some AMR determinants
- Isolates from the same species are difficult to separate through culturing process – trial and error
- It's difficult to know the isolate of clinical relevance – subtyping might be useful



Isolate ID	Expected species	Output from Quality Control pipeline						Downstream analysis		
		Bases (MB)	Reads	N50	Nr. of contigs	Longest	Total BPs	Species	MLST	Resistance
<b>Isolate J</b>	Staphylococcus aureus	151	1053706	135887	117	342404	5038817	Escherichia coli	ecoli[ST-155]	aadA1-like,blaOXA-1,floR-like,sul1,sul2
<b>Isolate J - reextract</b>	Staphylococcus aureus	145	1061786	105253	143	306865	5040601	Escherichia coli	ecoli[ST-155]	aadA1-like,blaOXA-1,floR-like,sul1,sul2

- It seemed like there was a contamination
- DNA was re-extracted and the results of the new sequence are a perfect match
- This indicates that in fact there was nothing wrong with the first sequencing attempt
- Likely there was misidentification of the expected species
- But it's also possible that the metadata of the isolates was swapped at some point

Isolate ID	Expected species	Output from Quality Control pipeline						Downstream analysis		
		Bases (MB)	Reads	N50	Nr. of contigs	Longest	Total BPs	Species	MLST	Resistance
<b>Isolate M</b>	Staphylococcus aureus	14	101884	1102	1501	20858	1545320	Staphylococcus aureus	saureus[Unknown ST]	blaZ
<b>Isolate M - rerun</b>	Staphylococcus aureus	313	2332424	103842	92	232550	4634362	Yersinia enterocolitica	versinia[ST-12]	vat(F)-like

- Re-running the same DNA yields very strange results that don't match previous results
- It's suspected that there was a mistake in selecting the DNA for re-run
- It's confirmed that one isolate included in the same batch has results in line with these unexpected results
- Re-run with the correct DNA shows results in line with first low quality results
- Re-running was enough (no need to re-extract)

Identification		Output from Quality Control pipeline					
ID	Expected species	Bases (MB)	Reads	N50	Number of contigs	Longest	Total BPs
Isolate G	Klebsiella pneumoniae	40	289816	2920	2346	29790	4994958
Isolate G_rerun	Klebsiella pneumoniae	335	2311772	375077	46	832865	5490232

## Acquired antimicrobial resistance gene - Results

Aminoglycoside						
No resistance genes found.						
Beta-lactam						
Resistance gene	Identity	Query/HSP	Contig	Position in contig	Phenotype	Accession no.
blaSHV-40	99.88	861/861	NODE_1116_length_1612_cov_0.739394	157..1017	Beta-lactam resistance	AF535128
Colistin						
No resistance genes found.						
Fluoroquinolone						
Resistance gene	Identity	Query/HSP	Contig	Position in contig	Phenotype	Accession no.
oqxB	98.95	3153/2286	NODE_754_length_2286_cov_0.534044	1..2286	Quinolone resistance	EU370913
oqxA	99.23	1176/1176	NODE_165_length_5002_cov_0.762462	3009..4184	Quinolone resistance	EU370913
Fosfomycin						
Resistance gene	Identity	Query/HSP	Contig	Position in contig	Phenotype	Accession no.
fosA	98.54	420/412	NODE_419_length_3254_cov_0.833067	2535..2946	Fosfomycin resistance	ACW001000079
Fusidic Acid						
No resistance genes found.						
Glycopeptide						
No resistance genes found.						
MLS - Macrolide, Lincosamide and Streptogramin B						
No resistance genes found.						

## Acquired antimicrobial resistance gene - Results

Aminoglycoside						
Resistance gene	Identity	Query/HSP	Contig	Position in contig	Phenotype	Accession no.
aph(6)-Id	100.00	837/837	NODE_31_length_4638_cov_1.636444	136..972		M28829
aph(3'')-Ib	100.00	804/804	NODE_31_length_4638_cov_1.636444	972..1775		AF321561
Beta-lactam						
Resistance gene	Identity	Query/HSP	Contig	Position in contig	Phenotype	Accession no.
blaSHV-40	99.88	861/861	NODE_2_length_669309_cov_4.564226	471090..471950	Beta-lactam resistance	AF535128
blaTEM-1B	100.00	861/861	NODE_35_length_2170_cov_1.549682	662..1522	Beta-lactam resistance	AY458016
Colistin						
No resistance genes found.						
Fluoroquinolone						
Resistance gene	Identity	Query/HSP	Contig	Position in contig	Phenotype	Accession no.
oqxB	98.95	3153/3153	NODE_4_length_437966_cov_5.772371	421397..424549	Quinolone resistance	EU370913
oqxA	99.23	1176/1176	NODE_4_length_437966_cov_5.772371	424573..425748	Quinolone resistance	EU370913
Fosfomycin						
Resistance gene	Identity	Query/HSP	Contig	Position in contig	Phenotype	Accession no.
fosA	98.54	420/412	NODE_5_length_375077_cov_5.547881	14046..14457	Fosfomycin resistance	ACW001000079
Fusidic Acid						
No resistance genes found.						
Glycopeptide						
No resistance genes found.						
MLS - Macrolide, Lincosamide and Streptogramin B						
Resistance gene	Identity	Query/HSP	Contig	Position in contig	Phenotype	Accession no.
mph(A)	100.00	906/906	NODE_29_length_6934_cov_2.188776	195..1100	Macrolide resistance	D16261

## In summary

- We need to be critical when analysing QC reports and look at several parameters
- We also need to consider the sequencing process as a whole (e.g. to detect cases of sample mix-up)
- We have to keep in mind that it is possible to obtain unexpected results (e.g. in cases where the preliminary species identification was wrong)
- Sequence quality will absolutely affect the downstream analysis - better to be safe than sorry

Jette Sejer Kjeldgaard

*jetk@food.dtu.dk*

# Exercise about bioinformatics tools for species identification and subtyping

**Technical training workshop #1**

**7-8 December 2022**

Jette Sejer Kjeldgaard, DTU

# **Exercises: tools for species identification and subtyping**

# Identification and characterisation of bacteria

Why do we need subtyping and characterisation?

To which level is subtyping needed?

# Species identification by WGS

In this exercise: two different approaches:

- kmerFinder
  - Full genome match to database
  
- rMLST
  - Typing based on the 53 genes encoding the bacterial ribosome protein subunits
    - (*rps* genes)
  
- Supplemented with data on serotype, MLST and cgMLST



# Species identification by WGS – different approaches

- **KmerFinder**

- Matches your sequence to database of published
  - By blasting small stretches (kmers) of query genome to databases
- Shows the most identical match – **genome wide** - plus additional matches
- Can show contamination
  - No threshold for ‘acceptable’ level of other species
    - Gene acquisition and annotation (other species)
- Can show polymicrobial infections directly from specimens
- Can be used to find similar bacteria as reference for SNP analysis
  - Accessible by accession # in match

# Kmerfinder – why was it developed?

## Abstract:

- WGS is becoming available as a routine tool for clinical microbiology
- If applied directly on clinical samples, this could further reduce diagnostic times and thereby improve control and treatment. A major bottleneck is the availability of fast and reliable bioinformatic tools
- This study was conducted to evaluate the applicability of WGS **directly on clinical samples** and to develop easy-to-use bioinformatic tools for the analysis of sequencing data.
- Thirty-five random **urine samples** from patients with suspected urinary tract infections were examined using **conventional microbiology, WGS of isolated bacteria, and direct sequencing on pellets from the urine samples**
- A rapid method for analyzing the sequence data was developed
- Bacteria were cultivated from 19 samples but in pure cultures from only 17 samples
- WGS improved the identification of the cultivated bacteria, and almost complete agreement was observed between phenotypic and predicted antimicrobial susceptibilities.

**Rapid whole-genome sequencing for detection and characterization of microorganisms directly from clinical samples** Hasman H, Saputra D, Sicheritz-Ponten T, Lund O, Svendsen CA, Frimodt-Møller N, Aarestrup FM.  
*J Clin Microbiol*, Jan;52(1):139-46, 2014.

# Kmerfinder – why was it developed? II

- Complete agreement was observed between
  - species identification,
  - multilocus sequence typing, and
  - phylogenetic relationships
- for ***Escherichia coli* and *Enterococcus faecalis*** isolates when the results of WGS of cultured isolates and urine samples were directly compared
- Sequencing directly from the urine enabled **bacterial identification in polymicrobial samples**
- Additional putative pathogenic strains were observed in some culture-negative samples
- WGS directly on clinical samples can provide clinically relevant information and **drastically reduce diagnostic times**
- This may prove very useful, but the need for data analysis is still a hurdle to clinical implementation. To overcome this problem, a publicly available bioinformatic tool was developed in this study

# Kmerfinder output

## KmerFinder-3.2 Server - Results

KmerFinder 3.2 results:

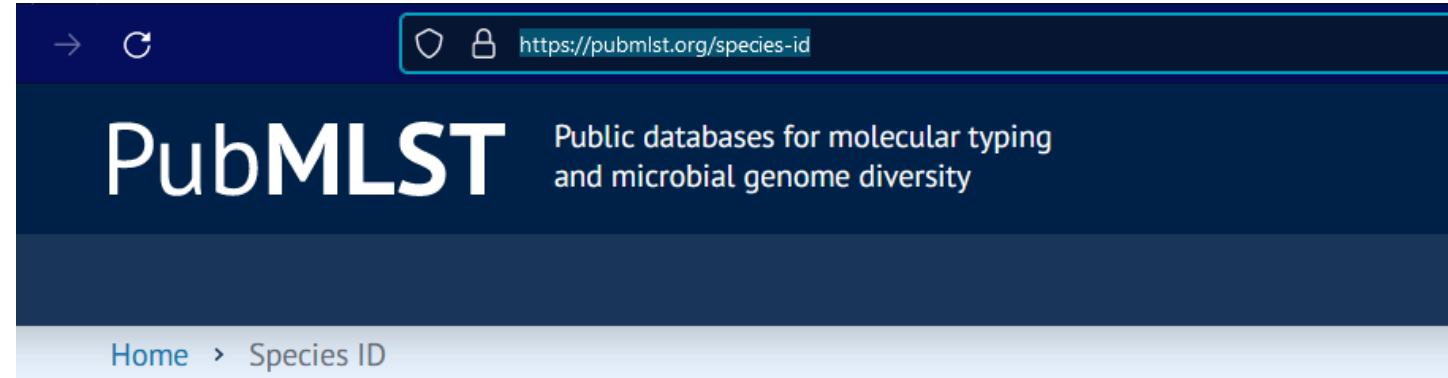
Template	Num	Score	Expected	Template_length	Query_Coverage	Template_Coverage	Depth	tot_query_Coverage	tot_template_Coverage	tot_depth
NZ_CP027586.1 Escherichia coli strain 2012EL-2448 chromosome, complete genome	9632	163881	3	166047	92.11	96.49	0.94	92.11	96.49	0.94
NZ_AP022362.1 Escherichia coli strain E302 chromosome, complete genome	12294	2012	69	168227	1.13	1.15	0.01	50.06	51.95	0.51
NZ_CP027340.1 Escherichia coli strain 2015C-3121 chromosome, complete genome	5590	1769	70	168958	0.99	1.02	0.01	90.14	92.74	0.90
NZ_CP015088.1 Escherichia coli O25b:H4 extrachomosomal sequence	25646	99	0	997	0.06	9.93	0.10	0.46	51.96	0.82

# Ribosomal MLST -rMLST

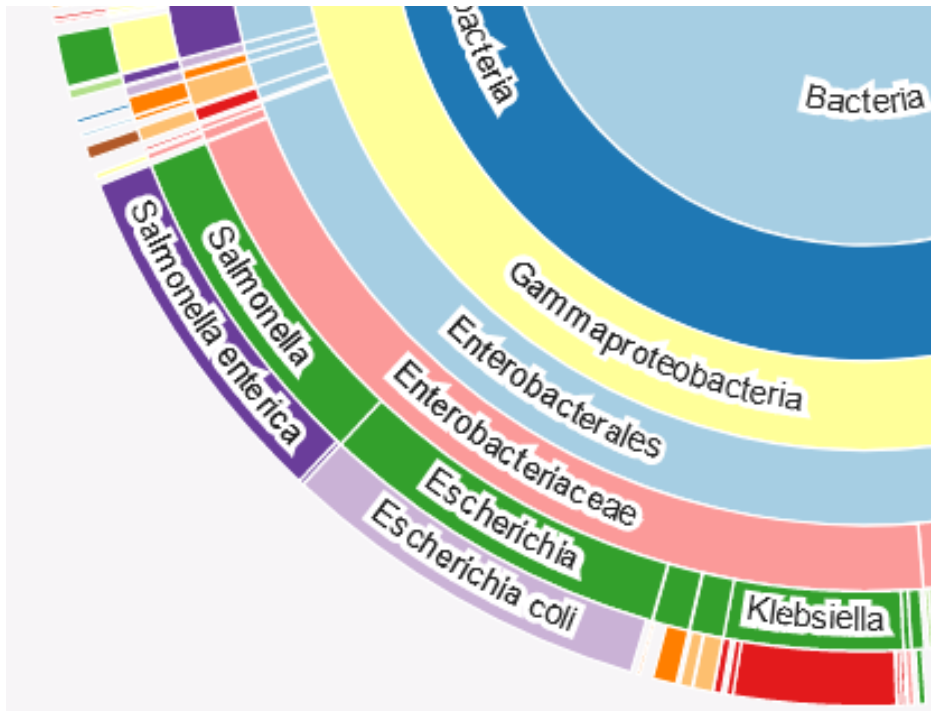
- Ribosomal Multilocus Sequence Typing (rMLST) is an approach that indexes variation of the 53 genes encoding the bacterial ribosome protein subunits (*rps* genes) as a means of integrating microbial taxonomy and typing.
- The *rps* gene variation catalogued in this database permits rapid identification of the phylogenetic position of any bacterial sequence at the domain, phylum, class, order, family, genus, species and strain levels.
- rMLST is described in [Jolley \*et al.\* 2012 \*Microbiology\* 158:1005-15](#)

<https://pubmlst.org/species-id>

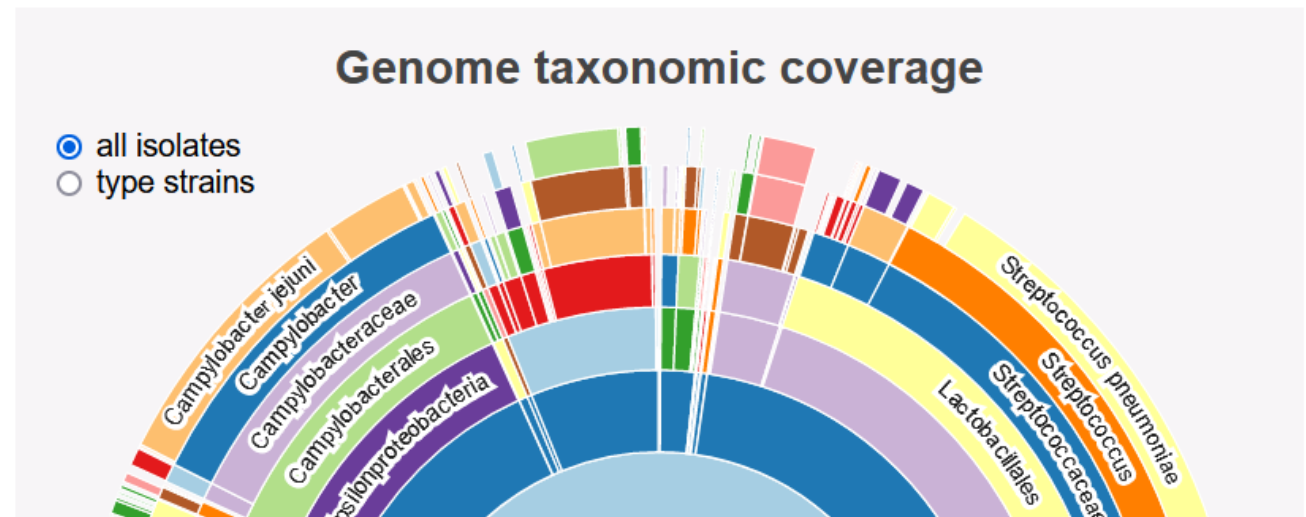
# rMLST



## Ribosomal MLST



IDENTIFY SPECIES



<https://pubmlst.org/species-id>



# rMLST output

Ribosomal MLST



Matching profile

rST: 1587

genus: Escherichia

species: Escherichia coli

## Predicted taxa

Rank	Taxon	Support	Taxonomy
SPECIES	Escherichia coli	100%	<i>Proteobacteria</i> > <i>Gammaproteobacteria</i> > <i>Enterobacterales</i> > <i>Enterobacteriaceae</i> > <i>E</i>

Uploaded file: EURGEN\_workshop\_isolate\_A.fa

55 exact matches found.

Locus	Allele	Length	Contig	Start position	End position	Linked data values	Flags
BACT000001 (rpsA)	1600	1674	NODE_1_length_181289_cov_8.918708	79214	80887	rMLST genome database species: <i>Escherichia coli</i> [n=3016]; <i>Escherichia sp.</i> [n=9]	
BACT000002 (rpsB)	27	726	NODE_31_length_61132_cov_9.793148	25554	26279	rMLST genome database species: <i>Escherichia coli</i> [n=10749]; <i>Shigella sonnei</i> [n=1813]; <i>Escherichia sp.</i> [n=42]; <i>Shigella sp.</i> [n=8]; <i>Shigella boydii</i> [n=3]	
BACT000003 (rpsC)	17	702	NODE_63_length_26316_cov_10.537401	5302	6003	rMLST genome database species: <i>Escherichia coli</i> [n=32127]; <i>Shigella boydii</i> [n=217]; <i>Escherichia fergusonii</i> [n=128]; <i>Escherichia sp.</i> [n=77]; <i>Shigella flexneri</i> [n=51]; <i>Shigella dysenteriae</i> [n=30]; <i>Shigella sp.</i> [n=16]; <i>Shigella sonnei</i> [n=14]; <i>Escherichia ruysiae</i> [n=6]	
BACT000004 (rpsD)	18	621	NODE_63_length_26316_cov_10.537401	13510	14130	rMLST genome database species: <i>Escherichia coli</i> [n=32411]; <i>Shigella sonnei</i> [n=672]; <i>Escherichia fergusonii</i> [n=148]; <i>Escherichia albertii</i> [n=119]; <i>Escherichia sp.</i> [n=94]; <i>Shigella sp.</i> [n=6]; <i>Shigella flexneri</i> [n=3]; <i>Escherichia whittamii</i> [n=2]; <i>Shigella boydii</i> [n=1]	
BACT000005 (rpsE)	28	504	NODE_63_length_26316_cov_10.537401	9956	10459	rMLST genome database species: <i>Escherichia coli</i> [n=13028]; <i>Shigella sonnei</i> [n=1797]; <i>Shigella boydii</i> [n=229]; <i>Shigella dysenteriae</i> [n=33]; <i>Shigella sp.</i> [n=25]; <i>Escherichia sp.</i> [n=16]	
BACT000006 (rpsF)	23	396	NODE_26_length_75429_cov_10.371902	72135	72530	rMLST genome database species: <i>Escherichia coli</i> [n=22306]; <i>Shigella sonnei</i> [n=1828]; <i>Shigella flexneri</i> [n=866]; <i>Shigella boydii</i> [n=234]; <i>Escherichia fergusonii</i> [n=151]; <i>Shigella sp.</i> [n=93]; <i>Escherichia sp.</i> [n=63]; <i>Shigella dysenteriae</i> [n=35]; <i>Escherichia albertii</i> [n=31]; <i>Escherichia ruysiae</i> [n=2]	
BACT000007 (rpsG)	18	471	NODE_4_length_156706_cov_9.663116	153231	153701	rMLST genome database species: <i>Escherichia coli</i> [n=31799]; <i>Shigella flexneri</i> [n=1093]; <i>Shigella sonnei</i> [n=250]; <i>Shigella boydii</i> [n=231]; <i>Escherichia albertii</i> [n=94]; <i>Shigella sp.</i> [n=83]; <i>Escherichia sp.</i> [n=57]; <i>Shigella dysenteriae</i> [n=35]; <i>Escherichia fergusonii</i> [n=10]	
BACT000008 (rpsH)	15	393	NODE_63_length_26316_cov_10.537401	8640	9032	rMLST genome database species: <i>Escherichia coli</i> [n=32422]; <i>Shigella sonnei</i> [n=1822]; <i>Shigella flexneri</i> [n=1130]; <i>Shigella boydii</i> [n=227]; <i>Escherichia fergusonii</i> [n=141]; <i>Escherichia sp.</i> [n=108]; <i>Shigella sp.</i> [n=93]; <i>Shigella dysenteriae</i> [n=35]; <i>Escherichia albertii</i> [n=1]	

# Additional data provided:

- MLST
  - Based on the seven-loci scheme #1
  - Widely used typing method for various bacteria
  - Can be used to determine intra-species contamination
    - Often no clear MLST type if there are several alleles of each locus
- cgMLST
  - Used for inferring phylogeny and show variation
  - Gives a nomenclature by cgST
- Serotyping
  - *E. coli* have traditionally been serotyped using antisera against the O-antigens and H-flagellar antigens
  - In the lab: generally time consuming and not always accurate
  - Supporting information of the WGS derived serotypes



## In this exercise:

- You will get outputs on 14 clinical *E. coli* isolates from blood or stool samples:
  - QC report
  - Species determination by kmerFinder and rMLST
  - cgMLST, rMLST and MLST – typing results
  - Serotypes
- Use the data to evaluate the sequence quality of isolates L, M and N, which seem problematic
- Discuss with your colleagues!
- It is not required to do your own analyses, but fasta files and direct links to analyses are available on Sciencedata or Slack, in case you want to look into the data yourself

# Materials:

- Two hardcopy sets of data, including some questions for consideration
- Optional: access to fasta files on sciencedata (link on Slack)
- Optional: access to direct links for analyses in the word file (on Slack)
- Optional: there are also links for cgMLST and SNP-based phylogeny of the remaining *E. coli* isolates (A-K)
  - We will discuss these tomorrow if we have time!

# Questions/considerations

- Questions to consider for L, M and N:
  - What seems to be wrong with this sequencing result?
  - How do you proceed with this isolate in the lab?
  - Can you use the sequence for e.g. resistance profiling?
- 
- General considerations:
  - Which methods can (preferably) be used to confirm/identify species and to identify contamination?
  - Discuss the discriminatory power of the different methods and compare which is more/less sensitive – and why?
  - How do we make cutoffs/thresholds for contamination with sequence analysis?

Bus leaves for the restaurant at 17:00!

Ana Rita Rebelo  
*anrire@food.dtu.dk*

# Questions and wrapping up the day

[EURGen-RefLabCap@food.dtu.dk](mailto:EURGen-RefLabCap@food.dtu.dk)

**Thank you on behalf of the  
EURGen-RefLabCap team**