E. coli virulence and pathotypes

Taxonomic analysis and typing

Dr. Jette Kjeldgaard (jetk@food.dtu.dk)
Technical University of Denmark (DTU), National Food Institute

DTU Food March 2026

Intended Learning Objectives

Specific objectives of this session:

1. Learn about taxonomic analysis and typing of *Escherichia coli*
2. Learn how to perform those analyses using WGS-based methods

Outline

This session consists of the following elements:

1. Introduction to taxonomic analysis and typing
2. Explanation of taxonomy of *E. coli*
3. General overview of methods for identification of *E. coli*
4. Detailed explanation of WGS-based methods for identification and typing of *E. coli*

Taxonomy of *E. coli*

Gram-negative, facultative anaerobic, motile and non-spore-forming rod-shaped coliform

Scientific classification of *E.coli* (taxonomic ranks):

Domain Bacteria → Kingdom Pseudomonadati → Phylum Pseudomonadota/Proteobacteria → Class Gammaproteobacteria → Order Enterobacterales → Family Enterobacteriaceae → Genus Escherichia

Originally named *Bacterium coli commune,* as a common inhabitant of the human gut by Theodor Escherich, German physician (ca. 1885).
He demonstrated that particular strains were responsible for infant diarrhea and gastroenteritis.

Re-classified and named after Escherich in the mid-20th century.

There are more than 700 different serotypes of *E. coli* distinguished by different surface proteins and polysaccharides

Taxonomy of *E. coli* – phenotypic methods

Early taxonomy was based on phenotypic traits:

- Gram-negative rod
- Lactose fermentation
- Indole formation
- Facultative anaerobic metabolism
- Normal inhabitant of the human gut

The serotyping era (1950-1980) O:H:K antigens
This allowed classification into approx. 700 serotype combinations
Examples: O157:H7 or O26:H11

O-antigen (LPS) ~180 types
H-antigen (flagella) ~50 types
K-antigen (capsule)

Serotyping was the first method to meaningfully discriminate pathogenic from commensal types e.g., O157:H7 being associated with severe haemorrhagic colitis

Serotype does not equal pathotype—many serotypes contain both harmless and virulent strains.

Taxonomy of *E. coli*

Virulence gene–based taxonomy (1980s–2000s)

Molecular biology advances allowed E. coli to be classified by virulence traits, not just serotypes
This gave rise to the **diarrheagenic E. coli pathotypes**, defined by gene content

Pathotype	Marker/gene(s)
EPEC	*eae* ± *bfp*
ETEC	*elt, est*
EAEC	*aggR, aaiC*
EIEC	*ipaH*
DAEC	*daaE, dra*
AIEC	phenotypic definition (adhesion/invasion)

Identification and typing of *E. coli*

Examples of methods for identification and typing of *E. coli* (excluding WGS)

→ Phenotypic methods

Selective media – MacConkey, Brilliance *E. coli*/coliform (chromogenic), TBX agar

API Strip 20E from Biomérieux - Commercial system with small-scale biochemical tests

Biotyping as mentioned before (Indole, lactose, glucose)

→ MALDI-TOF MS

Matrix-assisted laser desorption/ionization–time of flight mass spectrometry

→ PCR methods

Multiplex PCR (incl. pathotype detection)

Single-nucleotide polymorphisms (SNP) based identification (real time PCRs, multiplex PCRs)

PCR for other targets

Whole-genome sequence-based tools

WGS workflow – single isolates

Represent the real sequence of nucleotides (A-T-C-G) in the genome

Fasta files (or fastq)

```
>sample_1
ATGATGCAGCATACTTCTGTGTGGTACCG
GTCAGTCCGTTTGTTCTTGTGGCGAGTGT
>sample_2
TTTCTTGACCGCGACCGCCAATCTTACCT
AAAATCAGCCAAACCTATCCCATCGCGGA
```

A sample is collected → The bacteria are isolated → The DNA is extracted → The DNA is sequenced → The data are processed → We investigate the genomes

Whole-genome sequence-based tools

Identification by alignment against references

The whole genome will be aligned against the reference genomes in the database.

The tool will show which reference genome in the database is best match to our query genome.

Example: KmerFinder

Whole-genome sequence-based tools

KmerFinder for *E. coli* / *Shigella* → https://cge.food.dtu.dk/services/KmerFinder/

KmerFinder 3.2 results:

Template	Num	Score	Expected	Template_length	Query_Coverage	Template_Coverage	Depth	tot_query_Coverage
NZ_CP022457.1 Shigella sonnei strain 2015C-3566 chromosome, complete genome	21599	148514	3	145223	92.15	97.92	0.91	92.15
NZ_CP034060.1 Shigella flexneri strain FDAARGOS_535 chromosome, complete genome	7258	2419	51	137878	1.50	1.64	0.02	64.94
NZ_CP093400.1 Salmonella enterica subsp. enterica serovar Infantis strain R21.114	15559	1964	58	154470	1.22	1.25	0.01	5.45

EXTENDED OUTPUT

Input Files: *SAMEA3891664.fa*

RESULTS as text (tab separated) | Results as spa

Identification and typing of *E. coli*

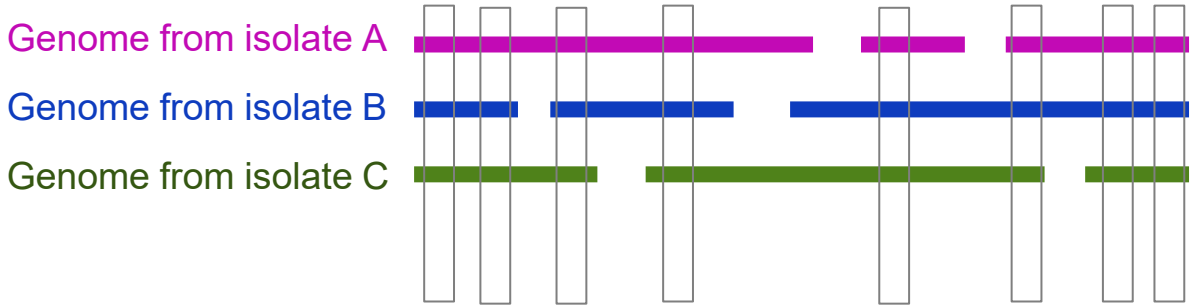
Typing with MLST

Multi-locus sequence typing

It is an **allelic profile scheme**.

It uses **conserved housekeeping genes** that are present in all isolates of the species.

For each bacterial species, the MLST scheme(s) contain specific and well defined genes (= loci).



Conserved loci that are present in all isolates from the species

Identification and typing of *E. coli*

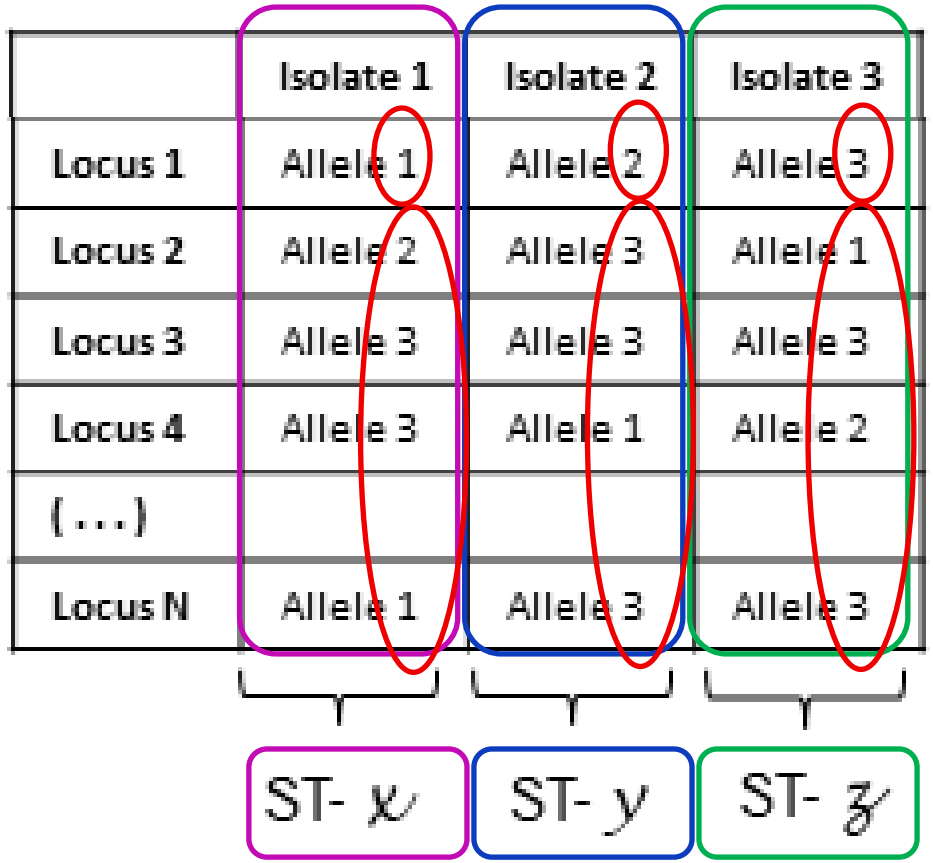
Typing with MLST

For each locus the specific genetic sequences can be different, because they accumulated mutations. Each specific sequence is called an allele.

Each bacterial isolate will have a **specific allele in each loci**.

The **exact combination** of alleles corresponds to a sequence type (ST).

If two isolates posses the same alleles they will belong to the same sequence type (ST). They are more **closely related** than isolates belonging to other STs.



Identification and typing of *E. coli*

MLST for *E. coli* → **https://enterobase.warwick.ac.uk/species/index/ecoli**

or https://cge.food.dtu.dk/services/MLST/

Two schemes developed for *E. coli:*

#1 Achtman (Enterobase and pubMLST)

#2 Pasteur (Hosted in the BIGSdb-Pasteur database)

Remember also, there were some Shigella-specific ST's

Both schemes used seven housekeeping genes – but

- they use different gene sets and therefore produce non-interchangeable ST numbers
- MLST STs do not directly define pathotypes, but some STs are strongly associated with certain pathotypes

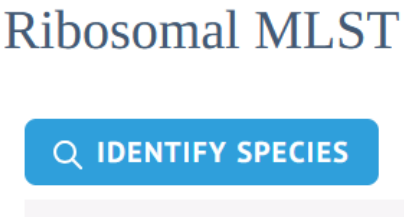
Examples: ST131 - globally dominant ExPEC clone (UTI, BSI),

ST38 – emerging carbapenem resistant clone,

ST11 – dominant pandemic O157:H7 clone, ST95 – leading NMEC lineage

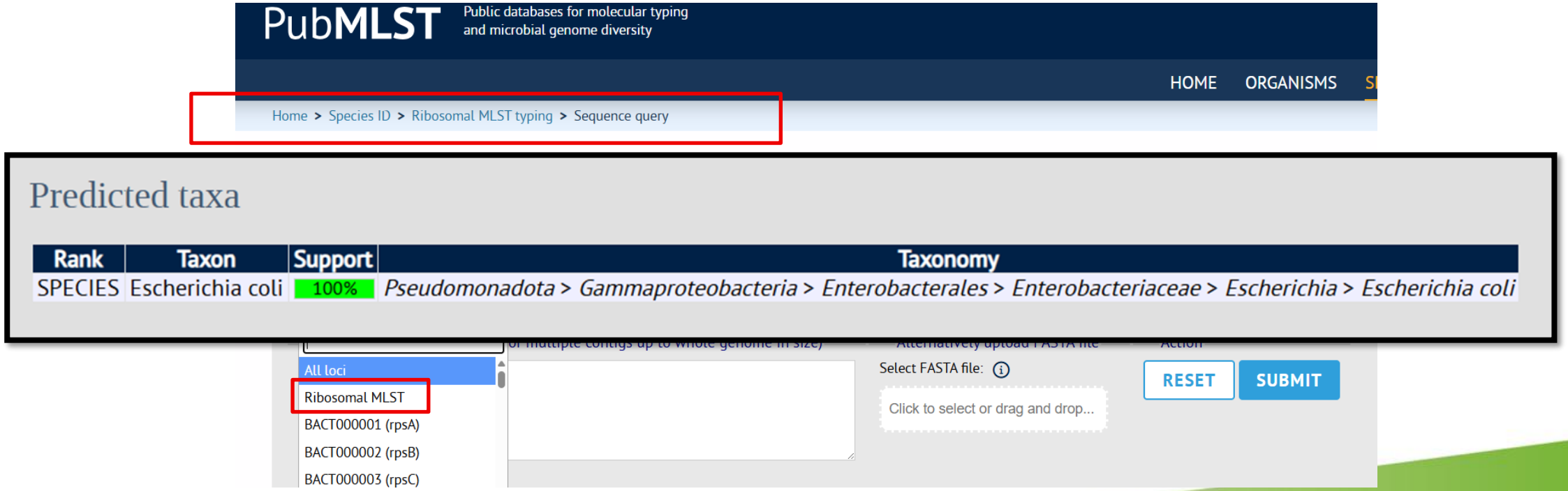
Whole-genome sequence-based tools

rMLST (ribosomal protein subunits) → https://pubmlst.org/species-id



Used for species identification.

Same principle as MLST but the target genes are different and the amount is higher (~53 loci).



Whole-genome sequence-based tools

Serotyping *in silico*: https://cge.food.dtu.dk/services/SerotypeFinder/

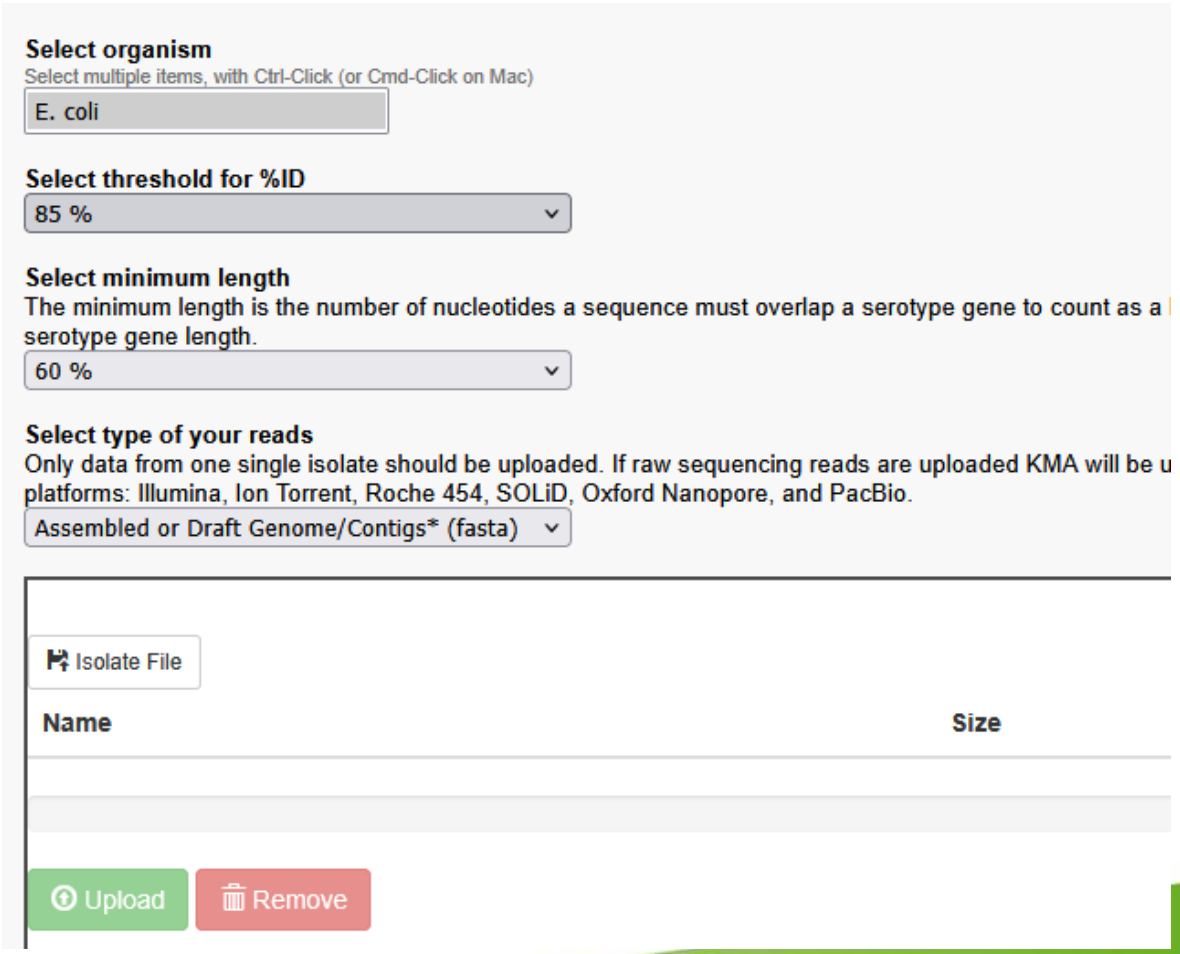
SerotypeFinder 2.0

SerotypeFinder identifies the serotype in total or partial sequenced isolates of E. coli.
Fasta file with test sequence: Test_sequence

Database(s): *O_type,H_type*

Database for O type genes						
Gene	Serotype	Identity	Template / HSP length	Contig	Position in contig	Accession number
wzx	O13/O129	99.84	1257 / 1257	SRR21910305_00030 len=38944 cov=30.8	18166..19422	AB972421
wzx	O13/O129	99.84	1257 / 1257	SRR21910305_00030 len=38944 cov=30.8	18166..19422	EU296422
wzy	O13/O129/ O135	99.56	1149 / 1149	SRR21910305_00030 len=38944 cov=30.8	15180..16328	AB972421
wzy	O13/O135	99.56	1149 / 1149	SRR21910305_00030 len=38944 cov=30.8	15180..16328	EU296422-EU296423

Database for H type genes						
Gene	Serotype	Identity	Template / HSP length	Contig	Position in contig	Accession number
fliC	H14	96.73	1653 / 1653	SRR21910305_00044 len=30684 cov=31.8	8306..9958	AY249998



Serotyping of *E. coli*

Serotyping has traditionally been used to identify presumptive EHEC/STEC strains

O157:H7 was originally the main type, but many more have been identified since, including non-motile variants (H-)

O26:H11(/H-); O103:H2, O111:H8(/H-), O121:H19, O45:H2, O145:H28…

Combining with ST lineages:
O26:H11 – major emerging STEC clone – ST21
O26 – hypervirulent lineage – ST29
O103:H2 - ST33
O111 – ST379

In summary

List of learning points in this session:

- *E. coli* comprises a very diverse group of bacteria, both intestinal commensals and pathogens
- Classically, *E. coli* has been typed into **serotypes**
- Phenotypic and molecular methods have been used to classify **pathotypes**
- Modern molecular methods incl. PCR and WGS allow for more strict differentiation and subtyping

- Two methods were shown to **identify** *E. coli* isolates:
  - **KmerFinder** (alignment-based)
  - **rMLST** (gene-by-gene approach with a specific MLST scheme)
- Two methods were shown to **subtype** *E. coli* isolates by:
  - Classic **MLST** (two schemes exist, with 7 genes each)
  - **Serotyping** – mapping of O- and H- antigens

Acknowledgements

The creation of this training material was commissioned by ECDC to Technical University of Denmark with the direct involvement of Jette Kjeldgaard. Some slides were adapted from Ana Rita Rebelo, DTU Food