Basic characterization: Serotyping, 7-genes Multi Locus Sequence Typing (MLST) and Virulotyping

Federica Gigliucci

Bioinformatics course,
19-20 October 2020

Istituto Superiore di Sanità, Dep. of Food Safety, Nutrition and Veterinary Public Health
European Union and National Reference Laboratory for *E. coli*, Rome, Italy

Serotyping

Serotyping, the 1st level of strain characterization

O : H

wzx, wzy, wzm, wzt ***fliC, flkA, fllA, flmA, flnA***

Strong evolutionary marker, it consents immediate detection of clinically relevant pathogens

NGS era!

Alignment (mapping or BLASTn) of genomic sequences VS database of reference genes sequences Joensen et al. JCM 2015

Galaxy / ARIES

E coli Serotyper Overview

This tool performs various operations:

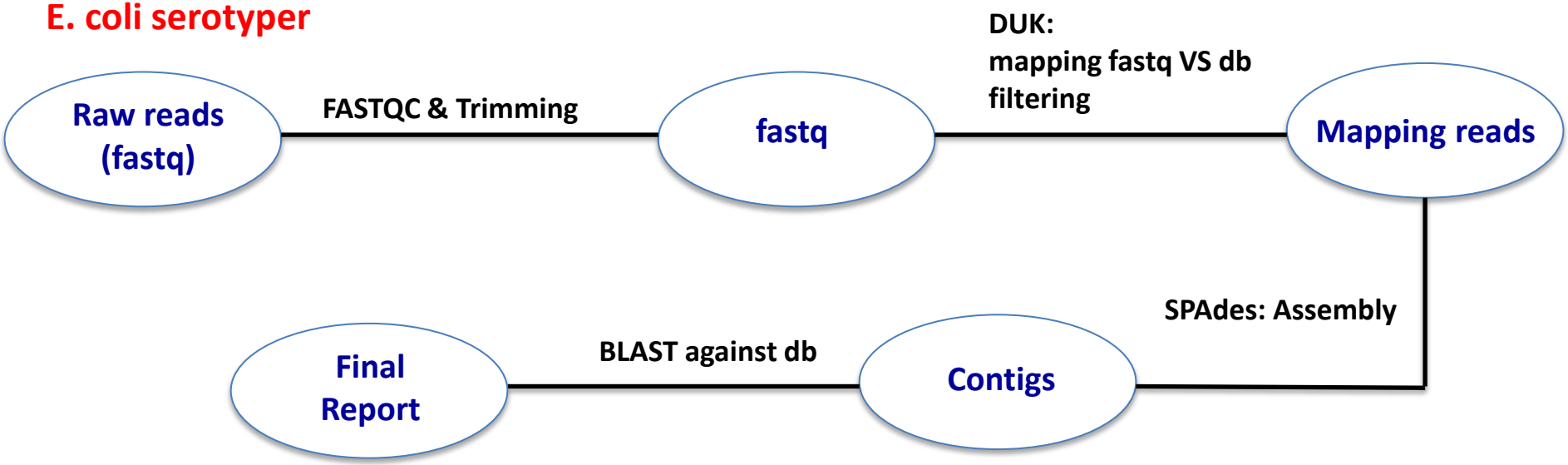
- Optionally: Quality assessment (FastQC)
- Optionally: Trimming (FASTQ positional and quality trimming)
- Optionally: Filtering (DUK)
- Optionally: Assembly (SPAdes)
- Serotyping (Blast+ against serotype databases from the Center for Genomic Epidemiology CGE)

Istituto Superiore di Sanità

European Union Reference Laboratory (EU-RL) for Escherichia coli, including Verotoxigenic E. coli (VTEC)

Developer: Arnold Knijn arnold.knijn@iss.it

E. coli serotyper



Serotyping - ARIES

Summary

O26:H11

Raw data quality check

FASTQC result forward: Webpage

FASTQC result reverse: Webpage

Best serotype match

FASTQC report, if the data analysed doesn't achieve minimum quality parameters O?:H11, O26:H?, O?:H?

Serotyping

sseqid	pident	length	positive
wzy_192_AF529080_O26	100.00	1023	1023
wzx_208_AF529080_O26	99.92	1263	1262
fliC_269_AY337465_H11	99.93	1459	1458
fliC_276_AY337472_H11	99.79	1459	1456

Choosing the best allele matching for each gene found (95% identity and with alignment length >800 bp)

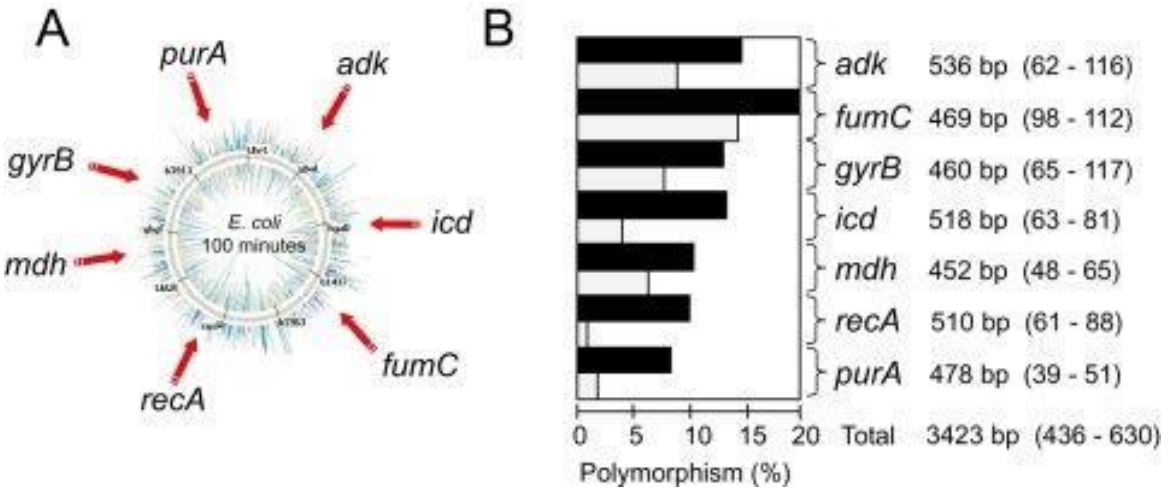
7-genes Multi Locus Sequence Typing (MLST)

Sequence Type (ST), the 2nd level of strain characterization

Deeper discriminatory power in case of outbreak investigation

MLST : Molecular typing of 7 house-keeping genes define the ST of bacterial strains

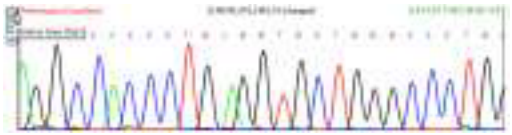
***E. coli* MLST scheme, by T. Wirth *et al*., Mol Microbiol 2006**



7-genes Multi Locus Sequence Typing (MLST)

Old era
conventional Sanger sequencing

PCR, sequencing, electropherograms analysis



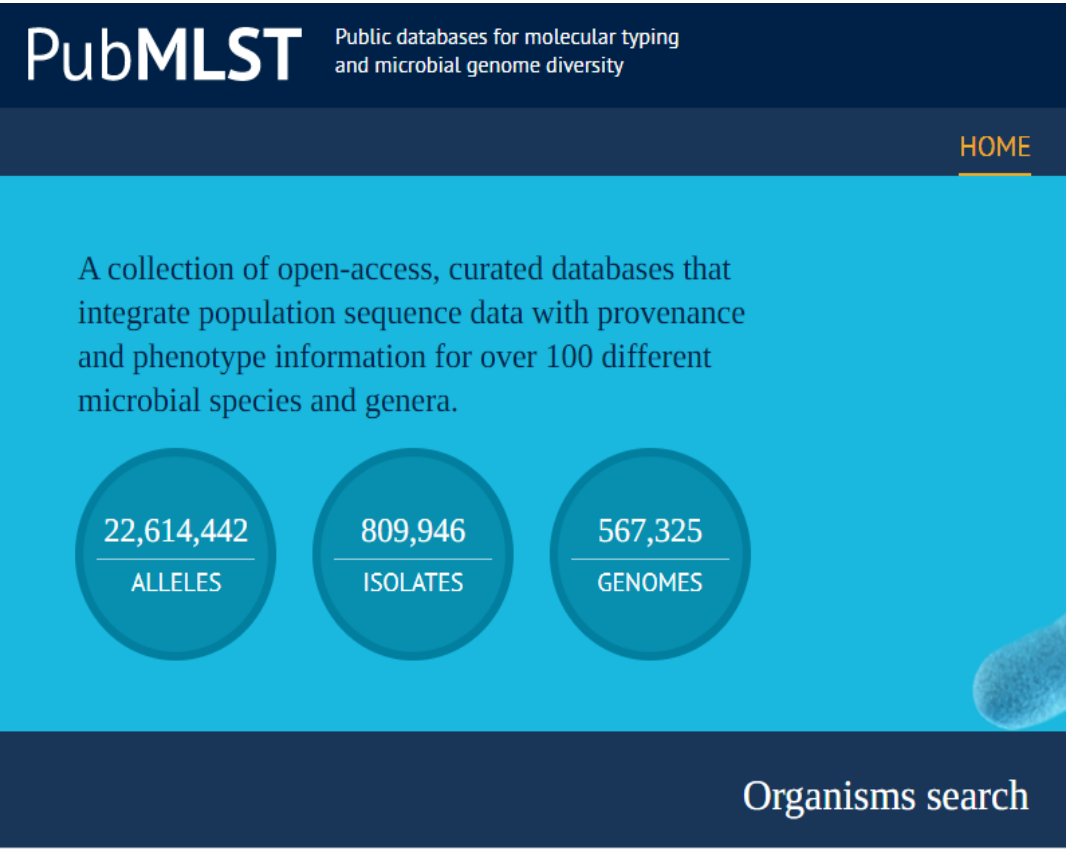
Uploading sequences on a webserver to obtain the corresponding alleles and STs

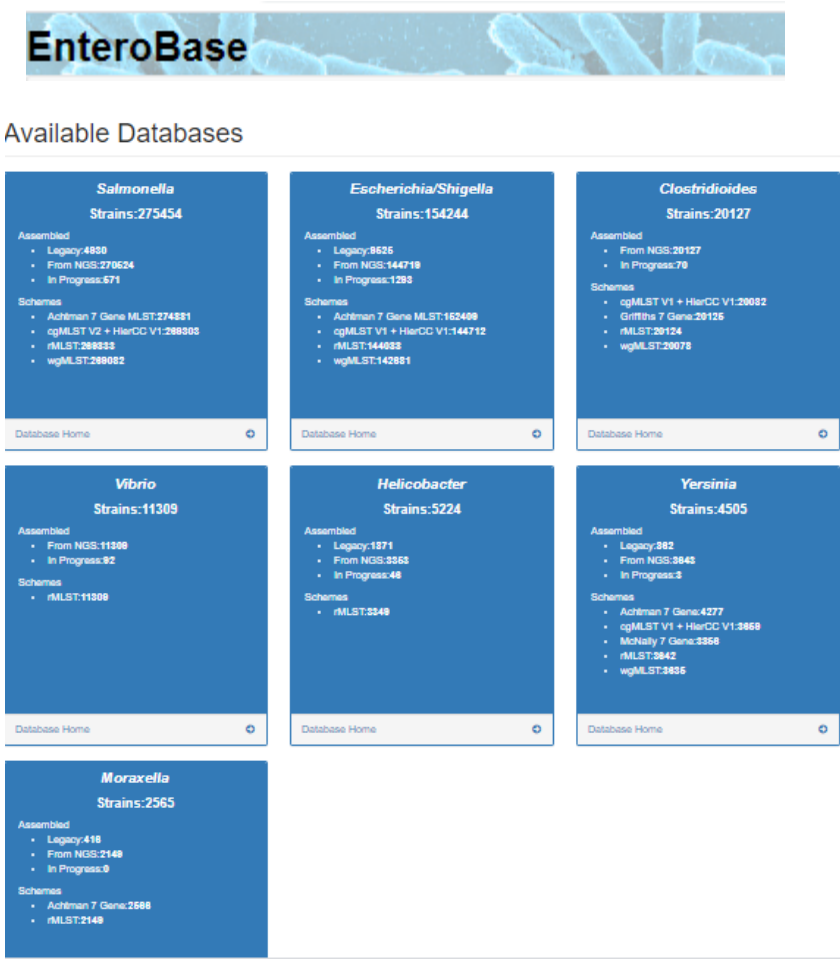
NGS era

Direct upload of WGS contigs on a webserver (e.g. ARIES)

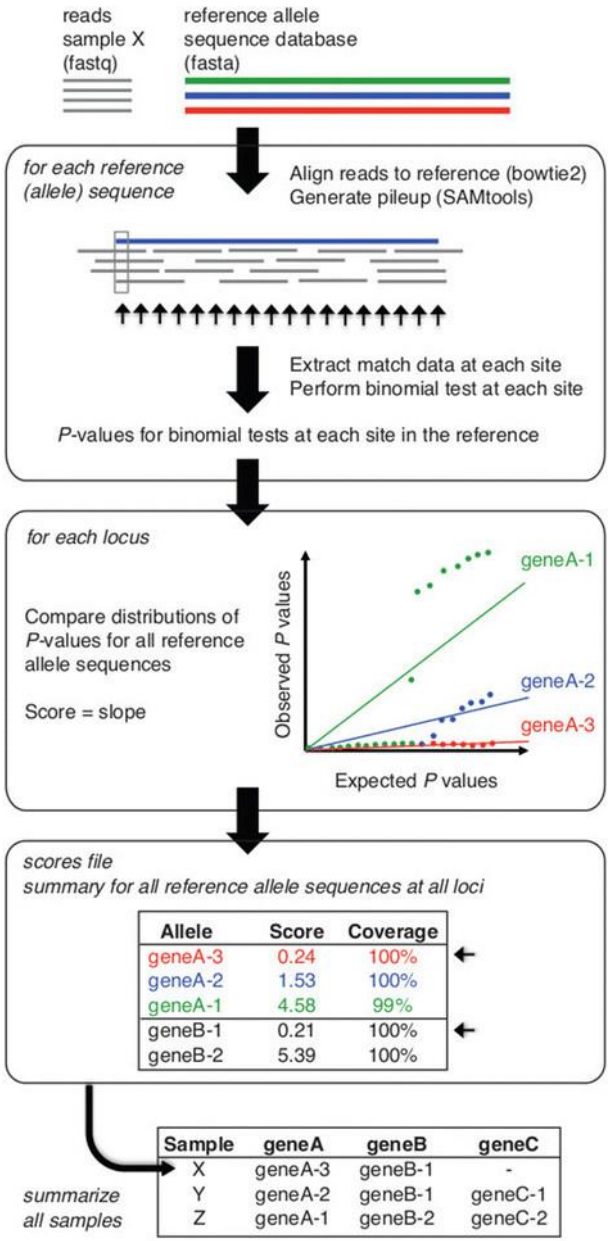
Alleles are directly retrieved through blastn comparison with pre-installed database of alleles from University of Warwick with pre-compiled pipelines

Public databases hosting MLST schemes





MLST- ARIES



SRST2

Read mapping-based tool, It derives the ST from reads

- Reads are aligned to all reference sequences (using bowtie2) and each alignment processed (using SAMtools).

- Statistical analysis: to determine which of all known reference alleles is most likely present at a given locus, the P value distributions for known alleles are compared. The slope of the fitted line is calculated and taken as the score for that allele.

- For each locus, the allele with the lowest score is accepted as the closest matching allele (small arrows) and reported in the output table.

Inouye M *et al*., Genome Medicine 2014 **6**:90

MLST- ARIES

Inouye M *et al*., Genome Medicine 2014 **6**:90

SRST2, output

1	2	3	4	5	6	7	8	9	10	11	12	13
Sample	ST	adk	fumC	gyrB	icd	mdh	purA	recA	mismatches	uncertainty	depth	maxMAF
readsall	17	6	4	3	17	7	7	6	0	-	139.33	0.141242937853

Depth coverage as indicator of the sequencing quality

* indicates mismatches
? indicates uncertainty due to low depth in some parts of the gene
- indicates the gene was not detected (> %coverage threshold, --min_coverage 90)

MLST- ARIES

MLST

T. Seemann, 2016. mlst **Github** https://github.com/tseemann/mlst

- **It scans contig files against traditional PubMLST typing schemes**

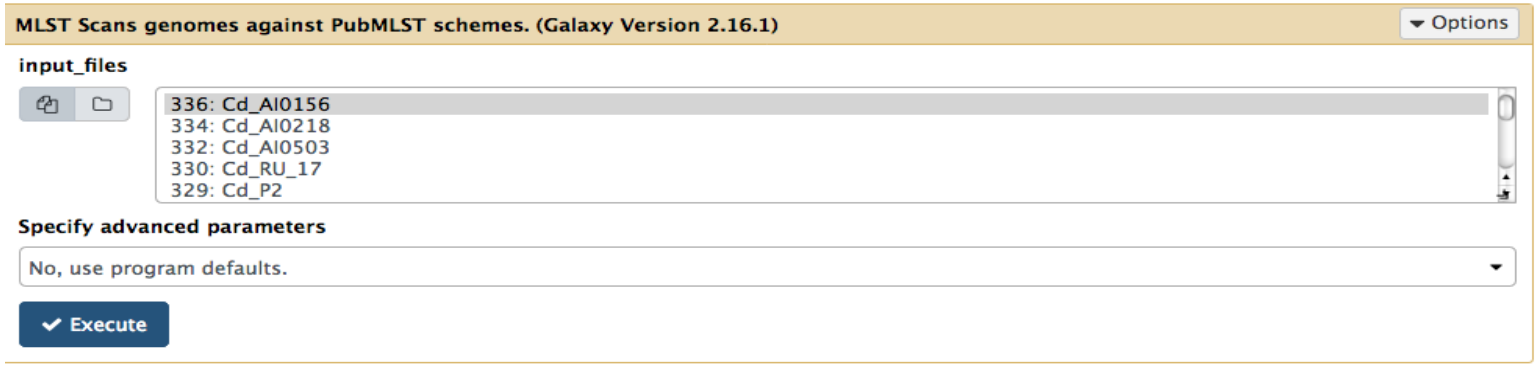
Available PubMLST schemes

abaumannii	bsubtilis	ecloacae	lmonocytogenes	pacnes	sgallolyticus	vcholerae
abaumannii_2	campylobacter	ecoli	lsalivarius	paeruginosa	shaemolyticus	vcholerae2
achromobacter	cbotulinum	ecoli_2	mabscessus	pdamselae	shominis	vibrio
aeromonas	cconcisus	edwardsiella	magalactiae	pfluorescens	sinorhizobium	vparahaemolyticus
aphagocytophilum	cdifficile	efaecalis	mbovis	pgingivalis	slugdunensis	vtapetis
arcobacter	cdiphtheriae	efaecium	mcanis	plarvae	smaltophilia	vvulnificus
bbacilliformis	cfetus	fpsychrophilum	mcaseolyticus	pmultocida_multihost	soralis	wolbachia
bcc	cfreundii	ganatis	mcatarrhalis	pmultocida_rirdc	spneumoniae	xfastidiosa
bcereus	chelveticus	hcinaedi	mhaemolytica	ppentosaceus	spseudintermedius	yersinia
bhampsonii	chlamydiales	hinfluenzae	mhyopneumoniae	pputida	spyogenes	ypseudotuberculosis
bhenselae	chyointestinalis	hparasuis	mhyorhinis	psalmonis	ssuis	yruckeri
bhyodysenteriae	cinsulaenigrae	hpylori	miowae	ranatipestifer	sthermophilus	
bintermedia	clanienae	hsuis	mmassiliense	rhodococcus	sthermophilus_2	
blicheniformis	clari	kaerogenes	mplutonius	sagalactiae	streptomyces	
bordetella	cmaltaromaticum	kkingae	mpneumoniae	saureus	suberis	
borrelia	cronobacter	koxytoca	msynoviae	sbsec	szooepidemicus	
bpilosicoli	csepticum	kpneumoniae	mycobacteria	scanis	taylorella	
bpseudomallei	csputorum	leptospira	neisseria	sdysgalactiae	tenacibaculum	
brachyspira	cupsaliensis	leptospira_2	orhinotracheale	senterica	tpallidum	
brucella	dnodosus	leptospira_3	otsutsugamushi	sepidermidis	ureaplasma	

MLST- ARIES

MLST

T. Seemann, 2016. mlst **Github** https://github.com/tseemann/mlst



- **It scans contig files against traditional PubMLST typing schemes**
- **It auto-detects bacterial species, just uploading the sequences**
- **Output: it produces a tab-separated file which contains: the filename - the closest PubMLST scheme name (bacterial species detected) - the ST - the allele IDs**

1	2	3	4	5	6	7	8	9	10
ED0257-phantastic_contigs.fasta	ecoli	11	adk(12)	fumC(12)	gyrB(8)	icd(12)	mdh(15)	purA(2)	recA(2)
ED1262-phantastic_contigs.fasta	ecoli	11	adk(12)	fumC(12)	gyrB(8)	icd(12)	mdh(15)	purA(2)	recA(2)
ED0597-phantastic_contigs.fasta	ecoli	11	adk(12)	fumC(12)	gyrB(8)	icd(12)	mdh(15)	purA(2)	recA(2)

Auto-detection good to find any possible contamination

MLST- ARIES

MLST

MLST does not just look for exact matches to full length alleles. It attempts to tell you as much as possible about what it found using the notation below:

Symbol	Meaning
n	Exact intact allele
~n	Novel full length allele similar to n
n?	Partial match to known allele
n,m	Multiple alleles
-	Allele missing

Setting **Output novel alleles** to true will produce an additional **novel_alleles.fasta** file containing the novel alleles.

Scoring system

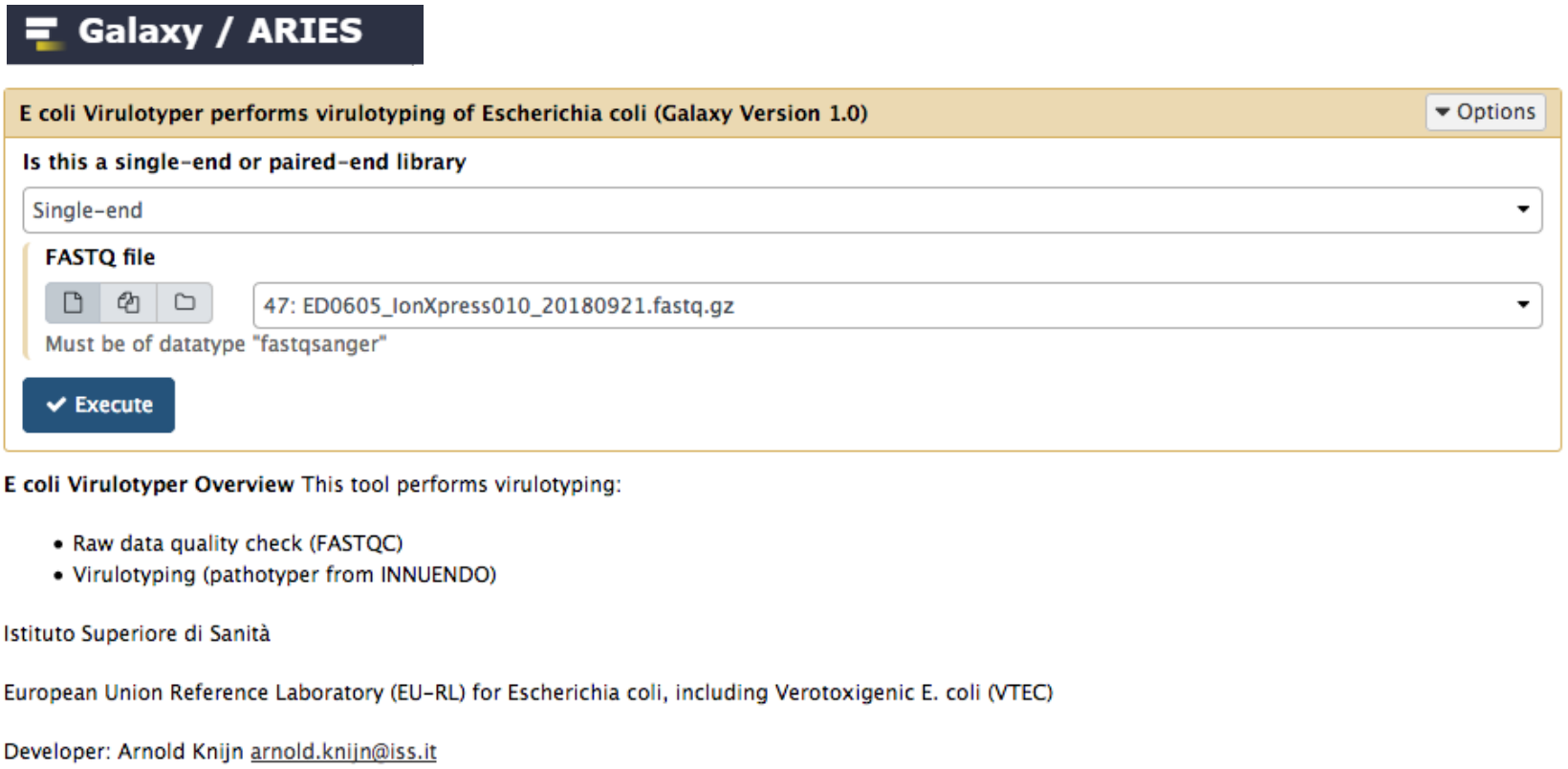
Each MLST prediction gets a score out of 100. The score for a scheme with N alleles is as follows:

+90/N points for an exact allele match e.g. 42
+63/N points for a novel allele match (50% of an exact allele) e.g. ~42
+18/N points for a partial allele match (20% of an exact alelle) e.g. 42?
0 points for a missing allele e.g. -
+10 points if there is a matching ST type for the allele combination

Virulotyping - ARIES

Virulence profile, the 3rd level of strain characterization

Do we have STEC strains?



Virulotyping - ARIES

- **Mapping (Bowtie2)** of the sequencing reads on the database
- Database of reference virulence genes sequences (in multiple allelic variants each) *E. coli* virulence finder database, Joensen JCM 2014
- Conversion of the output in a sam file (tabular) to extract interesting info and sequences
- Grouping of all the reads mapping to the different alleles for each gene
- Choosing the best allele matching for each gene found basing on the number of mapping reads and calculating the coverage
    - Percentage gene coverage (Gene length (min 90))
    - Gene mean read coverage (Gene depth coverage (min 15))
    - Percentage gene identity (min 90)

E coli Virulotyper

Report for Strain2_S5_L001_R1_001.fastq.gz

2019-06-26 10:35 UTC

Istituto Superiore di Sanità
Department of Food Safety, Nutrition and Veterinary Public Health
European Union Reference Laboratory for *E. coli*

Summary

eae, stx2A, stx2B

Raw data quality check

FASTQC result forward: Webpage

FASTQC result reverse: Webpage

Virulotyping

This table is filtered for results with >90% gene coverage, unfiltered results can be found here

#gene	percentage gene coverage	gene mean read coverage	percentage gene identity
espb_12_ecu65681	97.67	10.54	99.89
iss_13_cu928160	100.0	21.02	99.71
espb_13_af054421	97.57	11.39	99.67
nlec_6_ap010960	100.0	98.54	99.9
lpfa_3_ap010953	100.0	26.39	100.0
iss_11_ae014075	100.0	9.67	99.42
espa_22_fm201463	100.0	24.69	100.0
iss_7_cu928163	91.16	8.81	99.63
nlea_12_am422003	98.34	18.8	99.92
iss_8_cp001665	98.98	17.14	99.66
eae_45_ecu59503	97.66	36.98	99.89
prfb_13_cp002970	100.0	20.06	100.0
cif_2_ay128535	95.29	13.68	99.88
stx2b_27_ae005174_a	92.96	6.54	99.2
espj_1_ab303060	100.0	21.28	99.85
nleb_12_fm201463	92.93	12.39	99.89
nlec_3_ap010953	100.0	37.98	99.59
iss_12_cu928158	100.0	12.55	100.0

Best match for the main virulence genes associated with STEC

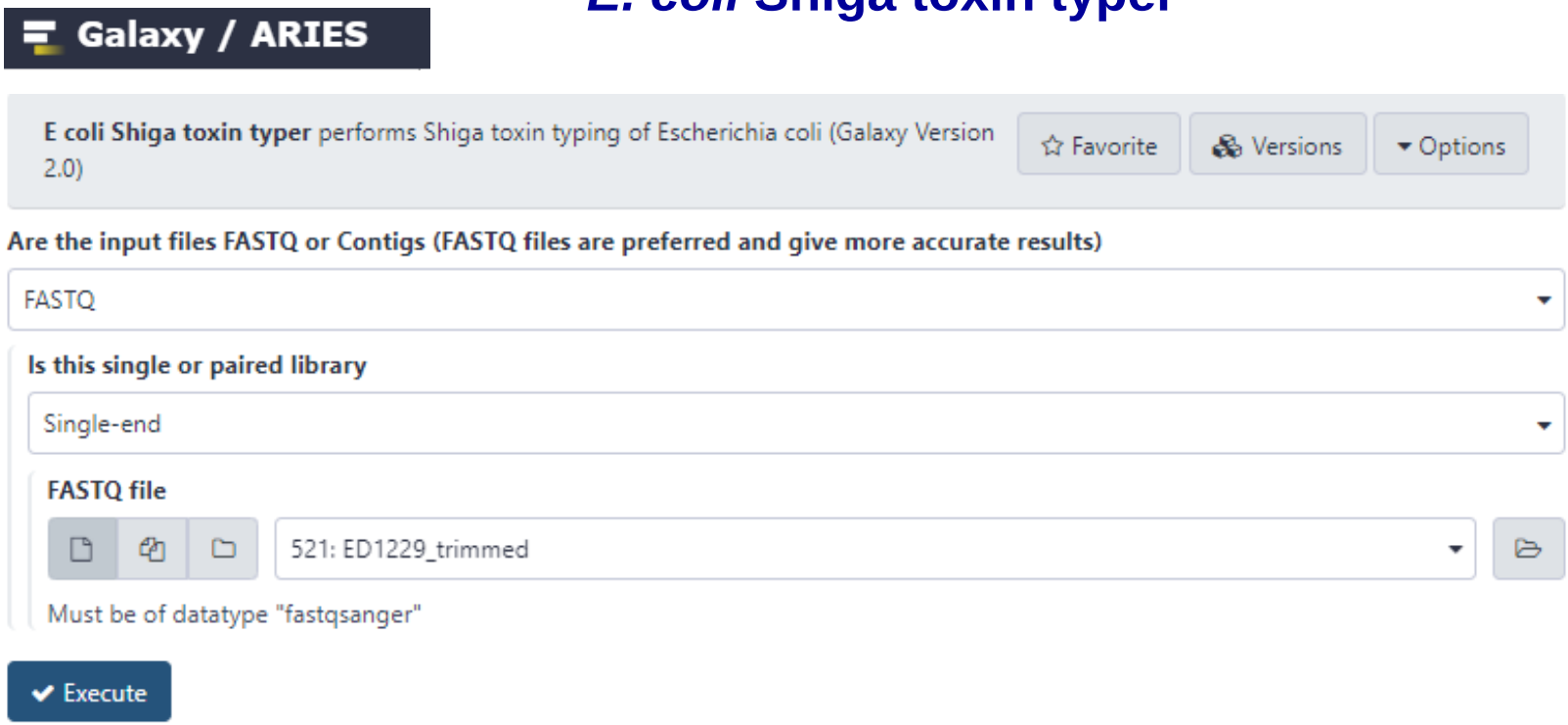
FASTQC report

Complete list of the best allele matching for each gene found

- Percentage gene coverage (Gene length (min90))
- Gene mean read coverage (Gene depth coverage (min15))
- Percentage gene identity (min90)

stx subtyping - ARIES

E. coli Shiga toxin typer



E coli Shigatoxintyper double-check version Overview

Comparison of the whole operon against the db

This tool performs various operations:

- Optionally: Quality assessment (FastQC v0.11.9)
- Optionally: Trimming (Trimmomatic v0.39)
- Optionally: Filtering (DUK v20110303)
- Optionally: Assembly (SPAdes v3.14 and SKESA v2.3)
- Optionally: Alignment (MUSCLE v3.8)
- Shigatoxintyping (Blast v2.9 against shiga toxin type databases from the Statens Serum Institut SSI and Technical University of Denmark DTU)

stx subtyping - ARIES

The tool accepts both raw reads (FASTQ) and contigs (FASTA)

Uploading contigs:

- Blastn search against the Shiga toxin subtype database (STSTDB) from the Statens Serum Institut SSI and Technical University of Denmark DTU (https://bitbucket.org/%7Bec84c234-a1e2-4442-8d73-bc3bdc479f29%7D/)

Uploading raw reads:

- FastQC and trimming of the raw reads
- Assembly and alignment of the contigs against the STSTDB, in order to construct *stx* consensus sequences on which the final blastn search will be performed
- Blastn search of *stx1* and *stx2* consensus sequences VS against the STSTDB, extracting the best matching sequence with an e-value < 0.001 and an identity > 95%.

stx subtyping - ARIES

E coli Shiga toxin typer

Report for

ED1246_IonXpress_040_20190723.fastq.gz

2020-10-12 08:18 UTC

Istituto Superiore di Sanità
Department of Food Safety, Nutrition and Veterinary Public Health
European Union Reference Laboratory for *E. coli*

Summary

stx2c

Best *stx* subtype match (95% < identity <= 100%)

Raw data quality check

FASTQC result: Webpage

FASTQC report

Shiga toxin typing

sseqid	pident	length	positive
stx2:52:AB015057:52	100.000	1241	1241

Choosing the best allele matching for each gene found (95% identity and with alignment length >800 bp)