Introduction to the EURL-VTEC WGS PT pipeline

Valeria Michelacci

Bioinformatics training,
July 2019

EURL-VTEC WGS PT pipeline

From the raw reads (.fastqsanger) to a complete report including:

- Quality check
- Trimming
- Serotyping
- 7 genes MLST
- virulotyping

Also returning trimmed reads files in output

Report for Ec201802549_S76_R1.fastq

2018-05-24 16:51 UTC

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

Summary

O26:H11
ST21

eae, ehxa, stx2a, stx2b

Best serotype match, ST and main virulence genes

(Disclaimer: The data analysed do not fulfill minimum quality parameters, please consider repeating the sequencing)

Raw data quality check

FASTQC result forward: Webpage

FASTQC result reverse: Webpage

If depth of coverage for 7 MLST genes is <30

Trimming analysis

Maximum length trimming	300
Left-side trimming	17
Right-side trimming	0
Minimum Phred quality score for right-side trimming	25
Average Phred quality score for right-side trimming	27
Minimum length filtering	-1

Serotyping – EURL-VTEC pipeline

Database of reference genes sequences **Joensen et al. JCM 2015**

O : H

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

Beta version:

alignment of the reads on the O and H databases

Assembly of the mapping reads

BLASTn match of the assembled contigs VS the serotype finder database

Serotyping

sseqid	pident	length	positive
wzy_192_AF529080_O26	99.90	1023	1022
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 MLST – EURL-VTEC pipeline

Mapping approach with SRST2 alignment tool

Reads are aligned on the MLST genes database of alleles

Returning alleles, ST, mismatches and depth of coverage

Multi Locus Sequence Typing

Sample	ST	adk	fumC	gyrB	icd	mdh	purA	recA	mismatches	uncertainty	depth	maxMAF
input	21	16	4	12	16	9	7	7	0	-	45.357	0.0869565217391

Virulotyping – EURL-VTEC pipeline

- Database of reference virulence genes sequences (in multiple allelic variants each) *E. coli* virulence finder database, Joensen JCM 2014
- **Alignment (Bowtie2)** of the sequencing reads on the database
- Conversion of the output in a sam file (tabular) to extract interesting info and sequences

QNAME	FLAG	RNAME	POS	MAPQ	CIGAR
ME2UT:01383:01267	0	gad:3:EF547388	1285	0	113M18I4M
ME2UT:02555:01592	16	gad:4:CP001925	1123	0	27M1I248M39I4M
ME2UT:02231:01820	0	gad:5:CP001846	87	1	138M
ME2UT:01605:00255	16	gad:5:CP001846	399	1	51M
ME2UT:01345:02031	16	gad:5:CP001846	685	1	176M
ME2UT:03330:02136	16	gad:5:CP001846	1050	1	6M1I38M
ME2UT:01475:02165	0	gad:6:BA000007	1	0	3M31I47M1D130M
ME2UT:01488:00709	16	gad:6:BA000007	1	0	4M32I55M1I149M
ME2UT:01943:01152	16	gad:6:BA000007	13	1	196M1I50M1I10M

- 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 mean read coverage
    - Percentage gene identity

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
ehxa_7_hm138194	95.43	15.16	99.83
nlea_8_ae005174	99.92	18.41	99.85
iss_13_cu928160	100.0	14.34	99.71
nlea_13_ap010960	93.35	10.76	99.84
katp_1_ab011549	100.0	80.34	100.0
iha_5_ap010953	100.0	41.51	100.0
espp_3_gq259888	100.0	31.96	99.95
nlec_6_ap010960	100.0	44.06	100.0
lpfa_3_ap010953	100.0	32.53	100.0
iss_7_cu928163	100.0	18.15	99.66
iss_8_cp001665	100.0	17.52	100.0
espp_4_ab011549	93.16	18.28	99.78
espp_1_hm138194	96.67	28.46	99.92
prfb_13_cp002970	100.0	27.98	100.0
cif_2_ay128535	99.88	13.41	100.0
espj_1_ab303060	100.0	15.39	100.0
iss_12_cu928158	100.0	19.68	98.25
stx2b_35_af525040_a	100.0	12.78	100.0
prfb_14_cp000800	97.39	21.73	99.77