Virulotyping and serotyping

Valeria Michelacci

NGS course, June 2016

Virulotyping – CGE-DTU

- Database of reference virulence genes sequences (in multiple allelic variants each) *E. coli* virulence finder database, Joensen et al. JCM 2014
- Accepting **contigs** or assembling input **reads** in contigs
- **BLASTn** match of the database of contigs VS the *E. coli* virulence finder database
- Choosing the best allele matching for each gene found (90% identity and covering a minimum of three-fifths of the length)

Virulence - E. coli						
Virulence factor	%Identity	Query/HSP length	Contig	Position in contig	Protein function	Accession number
stx2B	100.00	270 / 270	contig00123	1014..1283	Shiga toxin 2, subunit B, variant c	AB071845
nleB	100.00	981 / 981	contig00023	110..1090	Non-LEE encoded effector B	AE005174
nleC	100.00	993 / 993	contig00023	1151..2143	Non-LEE encoded effector C	AE005174
astA	91.96	112 / 117	contig00232	156..267	Heat-stable enterotoxin 1	AB042005
ehxA	99.97	2997 / 2997	contig00053	17315..20310	Enterohaemolysin	AB011549

Virulotyping - ARIES

- 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

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

- 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

1	2
virulence gene	coverage
astA:8:HM099897	14.359
cba:4:FJ664722	0.0729167
eae:42:AF071034	26.7522
efa1:6:AJ459584	1.32589
ehxA:3:AB011549	24.0854
epeA:1:AY258503	1.12549
espA:19:AE005174	12.1865

Serotyping – CGE-DTU

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

O : H

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

BLASTn match of the database of contigs VS the *serotype* finder database

Choosing the best allele matching for each gene found
(85% identity and covering a minimum of three-fifths of the length)

H type						
Serotype gene	%Identity	Query/HSP length	Contig	Position in contig	Predicted serotype	Accession number
fliC	99.82	1647 / 1647	out_39	43133..44779	H6	AIEY01000041

O type						
Serotype gene	%Identity	Query/HSP length	Contig	Position in contig	Predicted serotype	Accession number
wzy	99.47	1311 / 1311	out_46	7095..8405	O63	EU549862
wzx	99.92	1263 / 1263	out_46	9901..11163	O63	FJ539195

Predicted Serotype: O63:H6

Serotyping - ARIES

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

O : H

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

BLASTn match of the database of contigs VS the *serotype* finder database

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

1	2	3	4
wzx_208_AF529080_O26SSI	100.00	1263	1263
wzy_192_AF529080_O26SSI	100.00	1023	1023
wzy_191_DQ196413_O26	99.90	1023	1022
fliC_269_AY337465_H11	99.93	1459	1458
fliC_276_AY337472_H11	99.79	1459	1456