Assembly, assembly stats, virulotyping, serotyping

Valeria Michelacci

Bioinformatics training,
June 2018

Assembly

Short sequencing reads

.fastq file

```
@HWI-ST700693:238:B0224ACXX:1:1101:1218:1982
NACACTTGCTTTGGTGACAGCGGGGCATCCTCAAGC
+
#1=DDDDDHAFF?GEFGIIIIIIIIIIIIIIIIIFI
@HWI-ST700693:238:B0224ACXX:1:1101:1161:1986
NGATTTTGACCTCTCCAGTTTCCTCTTAACACTTTG
+
#1:BDFFFGHHHGJJJIIJHIJJJJJJJJJJJJJJJ
@HWI-ST700693:238:B0224ACXX:1:1101:1193:1989
NTATCCAGCCTGCGGTGCTACTTGGTGGAAGAGGAT
+
#1=DDFFFHGHGGJJFGHJJIJJJIEGECHDFHCC?
@HWI-ST700693:238:B0224ACXX:1:1101:1440:1981
NTCAAGAATCCAAGTGGGGCCAGCATAATGTACGCT
+
#1=DDFFFHGHDFDAEGIIFGIICGGHGBFGEFDHI
@HWI-ST700693:238:B0224ACXX:1:1101:1367:1983
NATTAGAACAGATCGCTACTTCGCCCGAAGATACAT
+
#4BDFFFFFHHHHHJGIJIJJJIJJJJJJJJJJJJIJ
@HWI-ST700693:238:B0224ACXX:1:1101:1395:1988
NTGGAAACGTTTTTAAACGCGGAGACAGCGTGGAGT
+
#1=DDFFFHCFFHJJJIJJIJJIJJJJGGIFHIGI7
@HWI-ST700693:238:B0224ACXX:1:1101:1285:1994
NCTTTGCTGTATTGACCGTTTGTAGATTTGAATCTT
+
#4=DDFFFHBHHHHIGIJFHIJFGGGIGIHIJIJII
@HWI-ST700693:238:B0224ACXX:1:1101:1632:1989
NTCTATGAATGTTCAAGCGGTAGCTGAGGAGAGTCC
+
```

FastqSize ≈ GenomeSize x Coverage x 2

At least 0.5 GB per genome

Partially assembled genome (contigs)

.fasta file

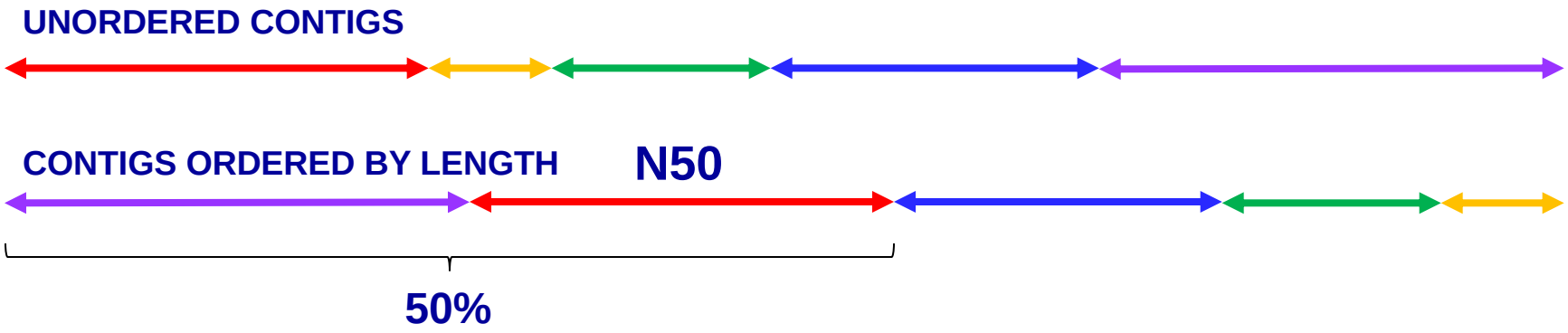
```
>NODE_1_length_449_cov_4.835189
ATCTTTCGCGCCTTCCAGCTCCAGCCATTCGGAACCGTTCGCCAGAAAACGGGCGTAATC
GGGTAAGACATAGCGCGGTTTGTACGGCGCATGACCTTCAAACATATCGCAGATTACACC
TTCATCCAGCGCGCGGCGGGCTTCGGCAGGAAGCTGTGGGTAAGGCAGATTGTTTTCTGC
TTCCAGTGCCAGAAAATGGCGCTTCTGCTCCGGGCTAAGCACTGGGCTGGTGACAATTTG
CTGGCAACGTTGTTGCAGTGCATTTTCATGAGAAGTGGGCATCTTCTTTTCCTTTTATGC
CGAAGGTGATGCGCCATTGTAAGAAGTTTCGTGATGTTCACTTTGATCCTGATGCGTTTG
CCACCACTGACGCATTCATTTGAAAGTGAATTATTTGAACCAGATCGCATTACAGTGATG
CAAACTTGTAAGTAGATTTCCTTAATTGTGATGTGTATCGAAGTGTGTTGCGG
>NODE_2_length_309_cov_4.686084
ACTGGTCAGTGCGGGTATCCTTGGACAATGGCCGATTGGACGTCTGGCGGATAAGTTTGG
TCGACTGCTGGTGTTGCGTGTTCAGGTCTTTGTCGTCATTCTCGGCAGTATCGCGATGCT
TAGCCAGGCGGCGATGGCCCCAGCGTTATTCATCCTCGGTGCCGCTGGCTTTACGCTATA
TCCGGTGGCGATGGCATGGGCTTGCGAGAAAGTTGAACATCATCAACTGGTGGCGATGAA
CCAGGCCTTACTGTTGAGCTATACTGTGGGAAGTCTGCTTGGCCCGTCATTTACCGCTAT
GCTAATGCAGAATTTCTCCGATAATTTATTGTT
>NODE_3_length_101_cov_3.346535
AGCGCATGAGCGCGCAGCGCCGCGTTACGTGGTGCATCAGCATGATGTTGGCCGGAGAG
TACAGAGACTCCCCTTCATCCATGATGCCCTCTTTCACCAGCAGTTCTTCAATCATCACC
AGACC
>NODE_4_length_311_cov_3.610933
CATCAACGCTAAAAGCCAGATGACGCAGACCGCAAGCTTCCGGTCGGCTGGGTCGTTCCG
GCGGGAACGGAAATGAGAAAAGCTCAATCACATATTGCCCATTAAGCGCCAAATCCCCTT
TCCATGAGTCGCGCGCTTCGCGATAGACTTCGCTTTGCAGCGTGAAACCAAGAATATCGC
AGTAGAAAGCTTTGCTCACCGCATAATCCGTCGCAATAATCGCAATATGGTGAACCTGTT
TTAAACCCAGCATAACGTCTCCTTTATTTGTTAACAGCACGTTACTCGCCCGGAAGCCGC
TCTGGCAAGTTATCCCGCCATTTTTAGGACTCGTA
>NODE_5_length_186_cov_4.973118
CGAAGATATAAGAAAGCGAACCAGAAAGAATGCCGGAGAACTTCATCAATTCATCACCTG
CATTGAGCAGATTTTGCAGGTTCTCAATAACCGGTAATCCAGCCCCAACGTTGGTGTCAT
AGAGGAATTTACGCCGCGATTTTTCCGCCGCATAACGCAACTGATGGTAGTAATCCATCG
ACGAGGTGTTGGCCTTTTTGTTCGGCGTGA
```

FastaSize for *E. coli* contigs

~5.5 MB

Assembly stats

N50 the **length** of the smallest contig among the set of the largest contigs that together cover at least 50% of the assembly



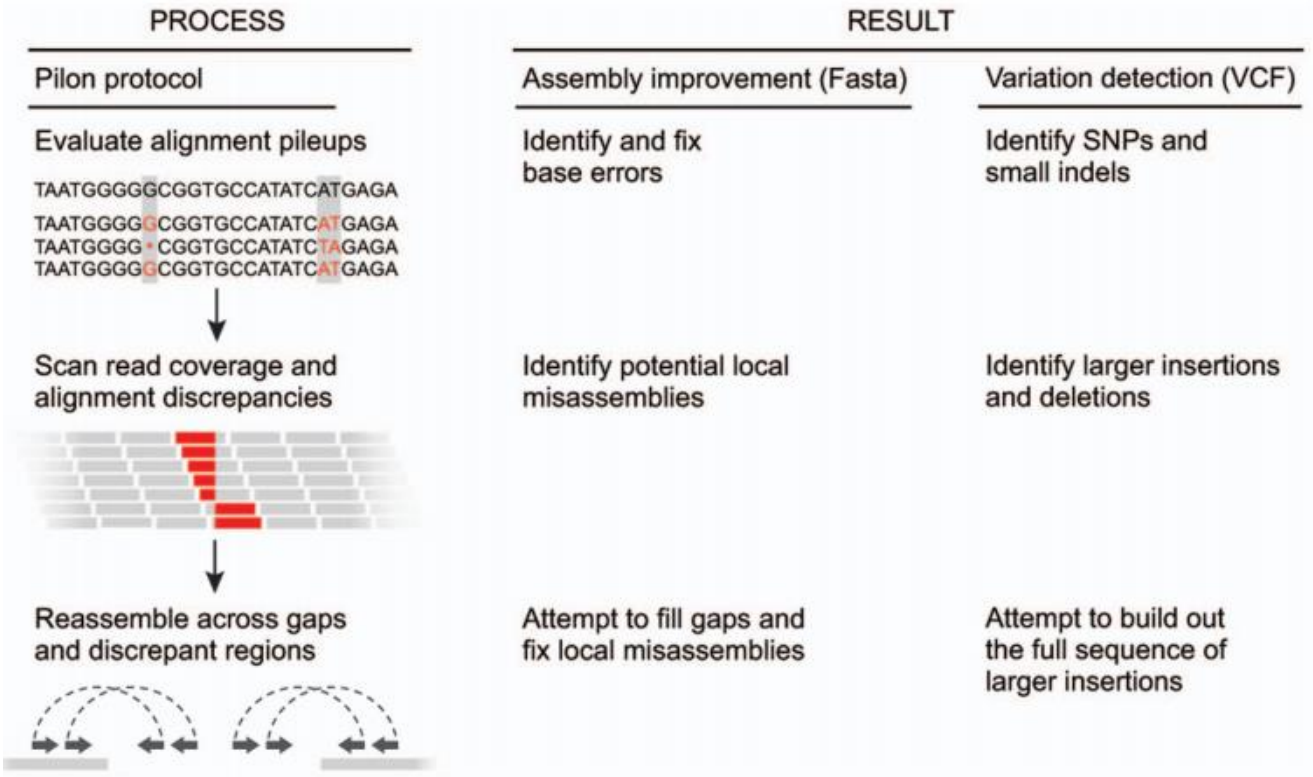
Other intuitive parameters to check:

Maximum contig length

Coverage of the contigs

Consensus length

Pilon – contigs refinement



Realignment of the reads on a «reference sequence»: we use Bowtie2 as alignment tool and the contigs as ref seq

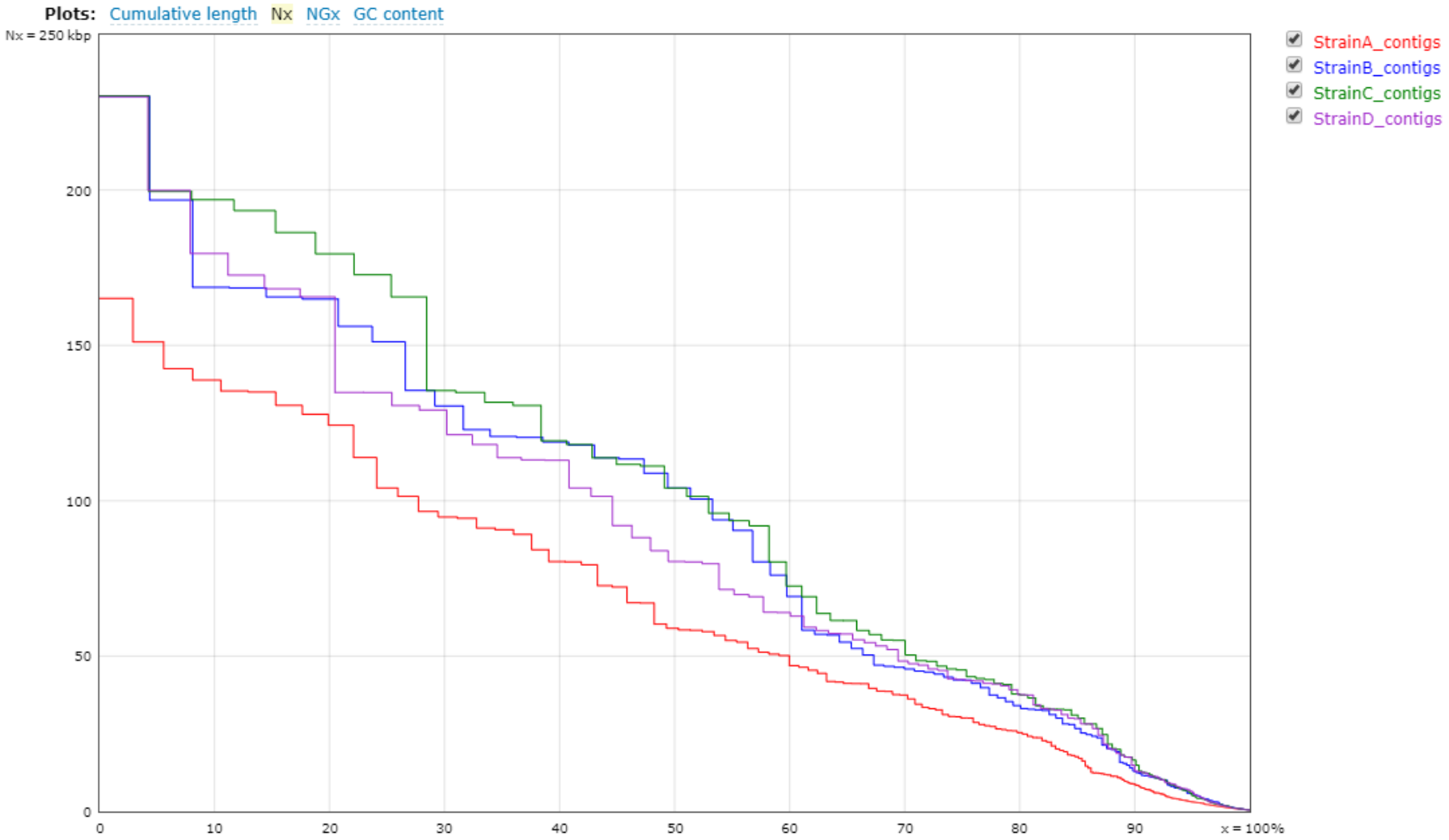
Pilon uses the result of the alignment to improve the assembly: it outputs better assembled contigs

Assembly stats: check bacterial contigs

Contigs Evaluator v1.0 on file dataset_126093.dat

Estimated genome size:	5000000 bp
Assembled nucleotides:	5754440 bp
Estimated coverage:	1.15 x
N. contigs:	818
Average contig length:	7035
Median contig length:	457
Maximum contig length:	165129
N. contigs >= 200 bp:	572 (69.9 %)
N. contigs >= 2,000 bp:	204 (24.9 %)
N50:	58429
NG50:	72678

Assembly stats: Quast



Blast searches

What we are used to:

online querying NCBI database for the presence of a sequence of interest

ONE SEQUENCE	**VS**	A DATABASE OF SEQUENCES (NCBI database)
	ONLINE (on NCBI webservers)	

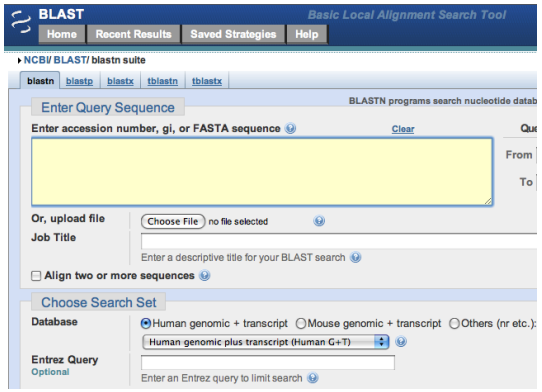
What we need now:

Inspect the contigs for the presence of interesting genes

ONE GENE	**VS**	A DATABASE OF SEQUENCES (OUR CONTIGS)
A DATABASE OF INTERESTING GENES	**VS**	A DATABASE OF SEQUENCES (OUR CONTIGS)

BLAST+ standalone suite

Possibility to install the blast+ suite locally to perform searches on custom databases



NCBI webserver

Command line operated tool

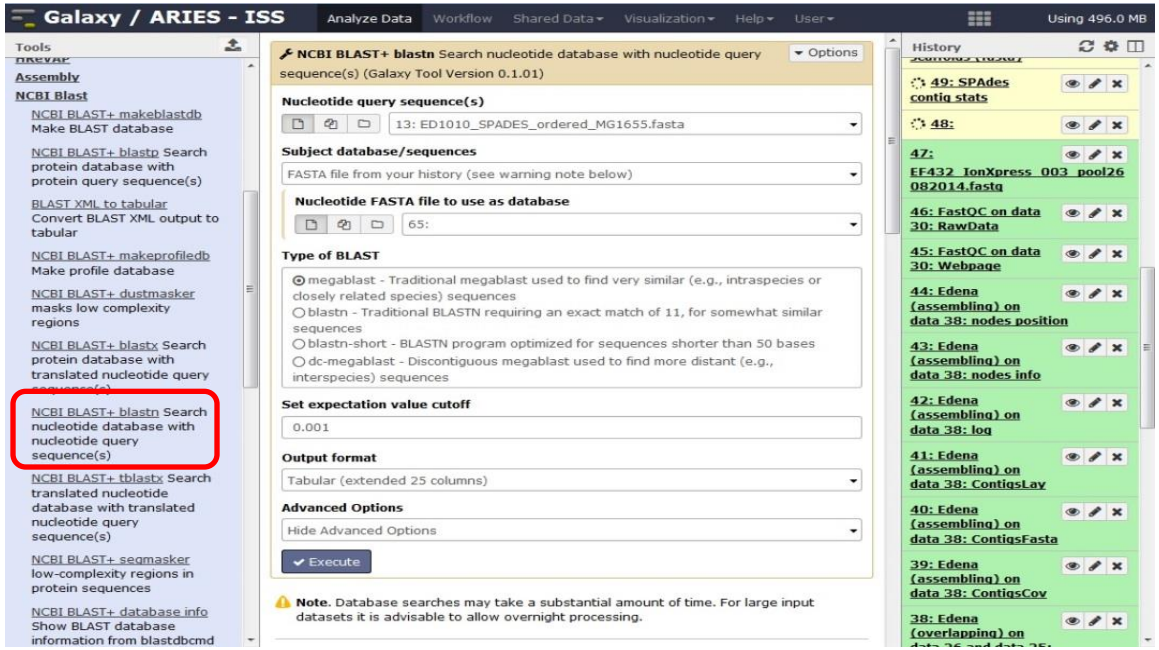
```
blastn -query text_query.txt -db refseq_rna.00 -out output.txt
```

This command instructs the system to:

- execute *blastn* program to search a nucleotide query against a nucleotide database
- use the sequence(s) in *test_query.txt* as the query
- search against the database *refseq_rna.00* database, and
- save the result in a file named *output.txt*

Available for Galaxy; currently running on **ARIES** (Galaxy @ISS)

To search for one gene: gene query VS database of contigs from the history



BLASTn output

Possibility to analyse the sequences for the presence of sequences of interest, compiled in custom databases **.fasta**

To search for a database of genes:
database query VS **database of contigs**

Downloadable output **.tab**

```
>gtrB_24111748
atgaaaatatctcttgtcgttcctgtcttcaatgaagaagaagcgatacctgttttctat
aaaacggtacgtgaattccaagagttgaagccatatgaagtagaaattgtattcataaat
gacggaagtaaagatgccacagagtcaattattaacgcgctggctgtttcagacccgcta
gttgttccgctgtcatttacacgcaacttttggtaaagaaccagccttatttgcagggtta
gaccatgcaagcggcgatgctgtaattcctattgatgtcgacctgcaagacccaattgag
gttatccctcatcttattgaaaagtggcaggcaggtgctgacatggtgcttgctaaacgt
tcagaccgctcaactgatggacgactgaaacgtaagacagctgagtggttctataaatta
cacaacaaaataagcaccccaaagatcgaggaaaatgtcggagattttcgactcatgtct
cgtgaggttgtggagaacattaaactgttgcctgagcgcaatcttttcatgaaaggcata
ctgagctgggtgggtggtcagacggatgtcgttgaatatgtacgcgcagagcgtgttgct
ggcatctcaaaatttaatggctggaaattatggaatctggcactggaaggtatcacaagt
ttttcaacctttcctcttcgcgtatggacttatataggcttgtttgttgcaagcatttca
ttttttatatggtgcatggatgattatagacacccttgtctttggtaacccagtacgcggg
tatccctccctgcttgtatcaatacttttcttgggtggagtgcaactgatcgggattggt
gttctcggagaatatataggtagaatctatattgaagtaaaaaatagacccaaatacatc
ataaaaaaatctcatcgaggtaacccatga
>gtrII_24111749
atgattaaaattaatttattaaaaaatgcaaatctccttgccttcatatcatgttttgca
atatcaatttattgttattgggggtggttgtacgatggaacacttaatattgatggcgag
tttacaaatatatttttatcaaacaatcacgcttgggcggtggtttcacacttttttgcga
cattactttctccctgagccttttttcactttatataacacccattaatagccttatctttt
atcatcatttcagcatttataatctgcagatcgctaaagctagaatcttatgaattattg
ataggtatgttagtatttattaccttccctcagatctcctatcaattagagtttcttaac
caagctgatactgtgggaattgcttttctactggcagcgatatcagcaattatttttcac
tcgcaaaaaaaataggattgtgatatttttctggtatagtactgtccattctttcaatggca
atctaccaaacattcgtaacatatattattgcattcgtcattgggttgcagataaattcg
ataatacgaaatgagaaaaatattcgtgaatctttttatagttcatgtttatctctatcc
ctcatagctttatctaccttaatttacctgctattaaccaaagctatcaagcattatttt
tcgcttgaatcgaacgagtacatctcaaattatatacaaaatgcaagcgatattaaatgg
cttgttaaatcagccatagataatatatataacttctataacaatcctcccactggttta
aacctatacaagtggttactgattcctttattaattctgatgtttaccctaacatataaa
ttaaaaacaagatcaatttatttgatttcatcaatcattttcatttatatacttccggtt
atatttatcgttgttgttggctcaggggcgccacctcgcctgtttgttttaatgcctata
gtagcagtaattttgttttcttgcttaagcaattttcgctctataaaatacctaaactgc
atgttttttttatttattatatatttaatggcgtttcaacatccaaaaatctatttttgaat
gatactctcgcaagacagaaagatatctctttagctaaagaaatatcatacacatctcaa
acaaaaggcatttcccttaacggaaaatatatatatatatatatggttcaaacgactcagga
aatatgctctccatgagcgcagacactttttggaaaatctttttttttggtgggatggtggc
aactatttttaggatggttgcatttatgaattactatggaatctgtaattgcaaaccagca
aataaagaacaaatagagaagatttatccaattgtgaagagtttaccttcctggccgaat
ccagattcgatagctgaaataaatggtttggtgataataaaactatcagagaaaaaaggg
tggcttccatttaatatttag
>iucA_24114944
ttgttgataatgagaatcattattgacataattgttattattttactgtgtgggagctgt
```

qseqid	sseqid	pident	length	mismatch	gapopen	qstart	qend	sstart	send	evalue	bitscore	sallseqid	score	nident	positive	gaps	ppos	qframe	sframe	qseq	sseq	qlen	slen
icsB_gi|18462582	contig00002	98.38	1485	24	0	1	1485	11764	13248	0.0	2571	contig000(	2850	1461	1461	0	98.38	1	1	CTATATAT	CTATATAT	1485	33481
ipaA_gi|18462581	contig00002	99.74	1903	4	1	1	1902	4378	6280	0.0	3406	contig000(	3776	1898	1898	1	99.74	1	1	TTAATCCT	TTAATCCT	1902	33481
ipaB_gi|18462580	contig00002	99.43	1743	9	1	1	1743	8448	10189	0.0	3095	contig000(	3432	1733	1733	1	99.43	1	1	TCAAGCA(	TCAAGCA(	1743	33481
ipaC_gi|18462579	contig00002	99.39	1149	7	0	1	1149	7337	8485	0.0	2040	contig000(	2262	1142	1142	0	99.39	1	1	TTAAGCT(	TTAAGCT(	1149	33481
ipaC_gi|18462579	contig00021	89.29	28	1	1	557	584	14639	14614	8,00E-04	35.6	contig000:	38	25	25	2	89.29	1	1	ATCCCTGA	ATCCCTGA	1149	18528
ipaC_gi|18462579	contig00209	100.00	19	0	0	221	239	3864	3846	8,00E-04	35.6	contig002(	38	19	19	0	100.00	1	1	TTACCAG(	TTACCAG(	1149	7232
ipaD_gi|18462578	contig00002	96.00	999	39	1	1	999	6289	7286	0.0	1618	contig000(	1794	959	959	1	96.00	1	1	TCAGAAA	TCAGAAA	999	33481
ipaH7.8_gi|18462574	contig00600	70.00	140	40	2	68	206	741	603	4,00E-10	57.2	contig006(	62	98	98	2	70.00	1	1	GTAATGA	GTAATGA	1698	2487
ipaH7.8_gi|18462574	contig00669	99.86	690	1	0	1	690	1384	2073	0.0	1240	contig006(	1374	689	689	0	99.86	1	1	ATGTTCTC	ATGTTCTC	1698	2073
ipaH7.8_gi|18462574	contig01015	99.86	705	1	0	816	1520	1	705	0.0	1267	contig010(	1404	704	704	0	99.86	1	1	CCCCCTG(	CCCCCTG(	1698	705

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

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
- 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

Virulotyping - ARIES

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