Decode NGS data: search for genetic features

Valeria Michelacci

NGS course, June 2015

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:

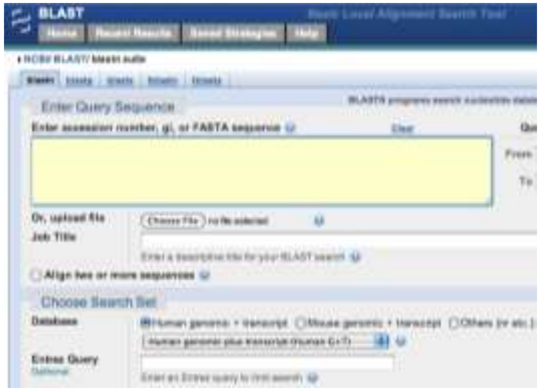
Inspect the contigs for the presence of interesting genes

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



BLAST Basic Local Alignment Search Tool

Home | Recent Results | Saved Strategies | Help

NCBI/ BLAST/ blastn suite

blastn | blastp | blastx | tblastn | tblastx

Enter Query Sequence

BLASTN programs search nucleotide databases

Enter accession number, gi, or FASTA sequence Clear

Query subrange From To

Or, upload file Choose File no file selected

Job Title

Enter a descriptive title for your BLAST search

□ Align two or more sequences

Choose Search Set

Database ● Human genomic + transcript ○ Mouse genomic + transcript ○ Others (nr etc.)

Human genomic plus transcript (Human G+T)

Entrez Query Optional

Enter an Entrez query to limit search

NCBI webserver

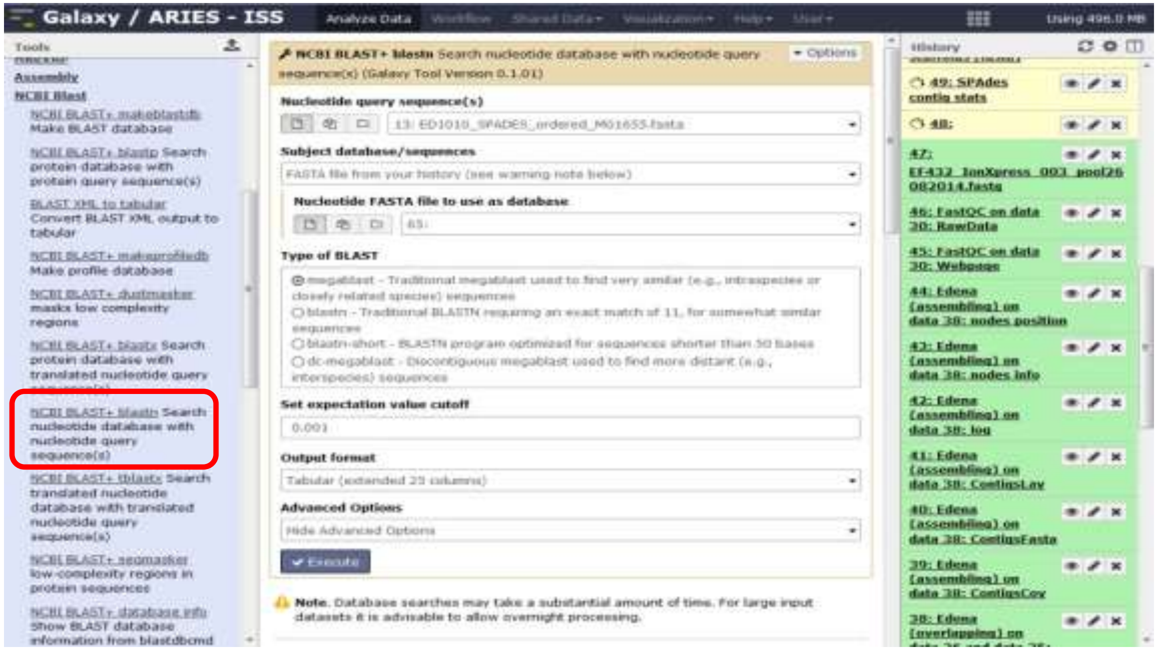
Command line operated tool

```
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)



Galaxy / ARIES - ISS
Analyze Data
Workflow
Shared Data
Visualization
Help
User
Using 496.0 MB
Tools
Assembly
NCBI Blast
NCBI BLAST+ makeblastdb Make BLAST database
NCBI BLAST+ blastp Search protein database with protein query sequence(s)
BLAST XML to tabular Convert BLAST XML output to tabular
NCBI BLAST+ makeprofiledb Make profile database
NCBI BLAST+ dustmasker masks low complexity regions
NCBI BLAST+ blastx Search protein database with translated nucleotide query sequence(s)
NCBI BLAST+ blastn Search nucleotide database with nucleotide query sequence(s)
NCBI BLAST+ tblastx Search translated nucleotide database with translated nucleotide query sequence(s)
NCBI BLAST+ segmasker low-complexity regions in protein sequences
NCBI BLAST+ database info Show BLAST database information from blastdbcmd
Options
History
49: SPAdes contig stats
48:
47: EF432_IonXpress_003_pool26 082014.fastq
46: FastQC on data 30: RawData
45: FastQC on data 30: Webpage
44: Edena (assembling) on data 38: nodes position
43: Edena (assembling) on data 38: nodes info
42: Edena (assembling) on data 38: log
41: Edena (assembling) on data 38: ContigsLay
40: Edena (assembling) on data 38: ContigsFasta
39: Edena (assembling) on data 38: ContigsCov
38: Edena (overlapping) on data 35 and data 35:
Execute

NCBI BLAST+ blastn Search nucleotide database with nucleotide query sequence(s) (Galaxy Tool Version 0.1.01)

Nucleotide query sequence(s): 13: ED1010_SPADES_ordered_M01655.fasta

Subject database/sequences: FASTA file from your history (see warning note below)

Nucleotide FASTA file to use as database: 65:

Type of BLAST

● megablast - Traditional megablast used to find very similar (e.g., intraspecies or closely related species) sequences
○ blastn - Traditional BLASTN requiring an exact match of 11, for somewhat similar sequences
○ blastn-short - BLASTN program optimized for sequences shorter than 50 bases
○ dc-megablast - Discontiguous megablast used to find more distant (e.g., interspecies) sequences

Set expectation value cutoff: 0.001

Output format: Tabular (extended 25 columns)

Advanced Options: Hide Advanced Options

Note. Database searches may take a substantial amount of time. For large input datasets it is advisable to allow overnight processing.

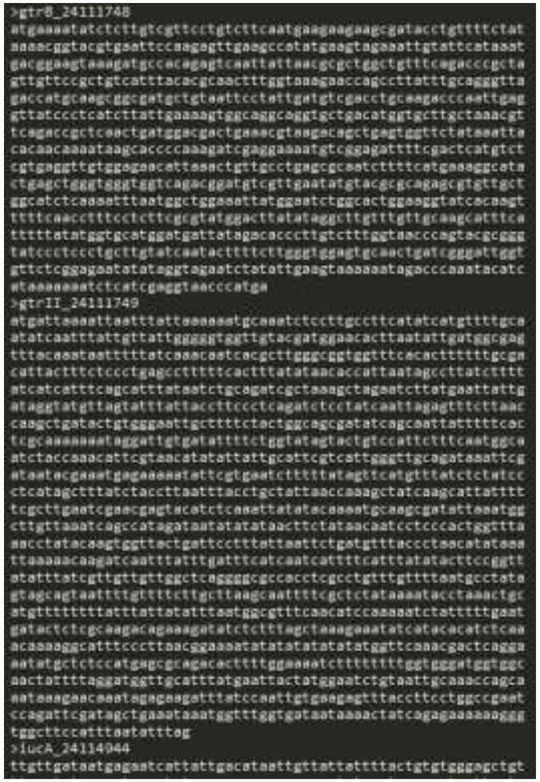
To search for one gene: gene query VS database of contigs from the history

BLASTn output

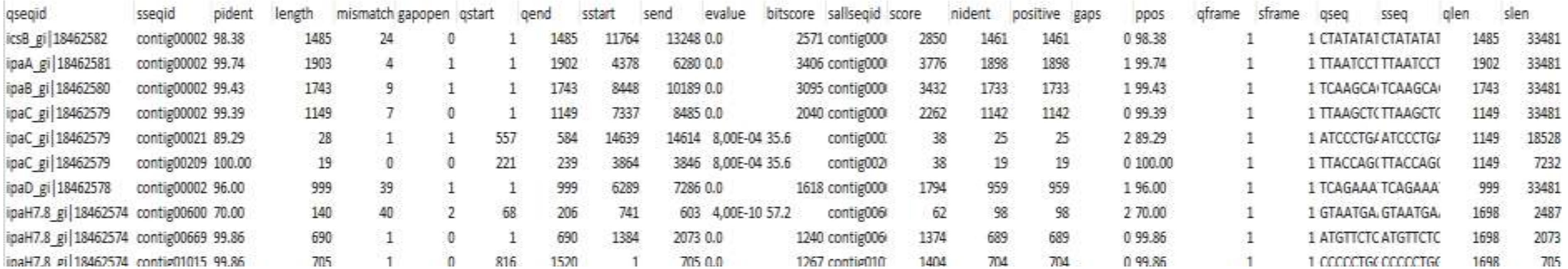
Possibility to analyse the sequences for the presence of sequences of interest, compiled in custom databases **.fasta**

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

Downloadable output **.tab**



```
>gtrB_24111748
atgaaaatatctcttgtcgttcctgtcttcaatgaagaagaagcgatacctgttttctat
aaaacggtacgtgaattccaagagttgaagccatatgaagtagaaattgtattcataaat
gacggaagtaaagatgccacagagtcaattattaacgcgctggctgtttcagacccgcta
gttgttccgctgtcatttacacgcaactttggtaaagaaccagccttatttgcagggtta
gaccatgcaagcggcgatgctgtaattcctattgatgtcgacctgcaagacccaattgag
gttatccctcatcttattgaaaagtggcaggcaggtgctgacatggtgcttgctaaacgt
tcagaccgctcaactgatggacgactgaaacgtaagacagctgagtggttctataaatta
cacaacaaaataagcaccccaaagatcgaggaaaatgtcggagattttcgactcatgtct
cgtgaggttgtggagaacattaaactgttgcctgagcgcaatcttttcatgaaaggcata
ctgagctgggtgggtggtcagacggatgtcgttgaatatgtacgcgcagagcgtgttgct
ggcatctcaaaatttaatggctggaaattatggaatctggcactggaaggtatcacaagt
ttttcaacctttcctcttcgcgtatggacttatataggcttgtttgttgcaagcatttca
tttttatatggtgcatggatgattatagacacccttgtctttggtaacccagtacgcggg
tatccctccctgcttgtatcaatacttttcttgggtggagtgcaactgatcgggattggt
gtttctcggagaatatataggtagaatctatattgaagtaaaaaatagacccaaatacatc
ataaaaaaatctcatcgaggtaacccatga
>gtrII_24111749
atgattaaaattaatttattaaaaaatgcaaatctccttgccttcatatcatgttttgca
atatcaatttattgttattgggggtggttgtacgatggaacacttaatattgatggcgag
tttacaaataattttatcaaacaatcacgcttgggcggtggtttcacactttttttgcga
cattactttctccctgagcctttttcactttatataacaccattaatagccttatctttt
atcatcatttcagcatttataatctgcagatcgctaaagctagaatcttatgaattattg
ataggtatgttagtatttattaccttccctcagatctcctatcaattagagtttcttaac
caagctgatactgtgggaattgcttttctactggcagcgatatcagcaattatttttcac
tcgcaaaaaaataggattgtgatattttctggtatagtactgtccattctttcaatggca
atctaccaaacattcgtaacatatattattgcattcgtcattgggttgcagataaattcg
ataatacgaaatgagaaaaatattcgtgaatctttttatagttcatgtttatctctatcc
ctcatagctttatctaccttaatttacctgctattaaccaaagctatcaagcattatttt
tcgcttgaatcgaacgagtacatctcaaattatatacaaaatgcaagcgatattaaatgg
cttgttaaatcagccatagataatatatataacttctataacaatcctcccactggttta
aacctatacaagtggttactgattcctttattaattctgatgtttaccctaacatataaa
ttaaaaacaagatcaatttatttgatttcatcaatcattttcatttatatacttccggtt
atatttatcgttgttgttggctcaggggcgccacctcgcctgtttgttttaatgcctata
gtagcagtaattttgtttcttgcttaagcaattttcgctctataaaatacctaaactgc
atgtttttttttatttattatatttaatggcgtttcaacatccaaaaatctatttttgaat
gatactctcgcaagacagaaagatatctctttagctaaagaaatatcatacacatctcaa
acaaaaggcatttcccttaacggaaaatatatatatatatatggttcaaacgactcagga
aatatgctctccatgagcgcagacacttttggaaaatcttttttttggtgggatggtggc
aactattttaggatggttgcatttatgaattactatggaatctgtaattgcaaaccagca
aataaagaacaaatagagaagatttatccaattgtgaagagtttaccttcctggccgaat
ccagattcgatagctgaaataaatggtttggtgataataaaactatcagagaaaaaaggg
tggcttcccatttaatatttag
>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	TTAAGCTC	TTAAGCTC	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	TTACCAGC	TTACCAGC	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	CCCCCTGC	CCCCCTGC	1698	705

Databases for *E. coli*

Need for reference databases!

Some are available, some still need curation

- **Serotype-associated genes**

E. coli O- and H- antigens databases available from SSI database (Joensen et al., JCM 2015)

- **MLST alleles**

Database of the alleles from the 7 housekeeping genes available from the University of Warwick website

- **Virulence genes**

VTEC virulence genes available from SSI database (Joensen et al, JCM 2014); work in progress @ISS for other pathogenic _E. coli_

Blastn-based tools available on DTU-CGE webservers and on ARIES Galaxy@ISS

Inspect for genetic features on ARIES

By selecting parameters and thresholds, the result can directly point at the genetic features searched most probably encoded on the strain sequenced

Easy to use pre-compiled pipelines available on ARIES:

E. coli Virulotyper

E. coli Serotyper

E. coli MLST Warwick

Moreover ARIES as any Galaxy is open and easy to use for analysing the presence of other genes of interest

Customizable – implementable
Useful for surveillance AND research

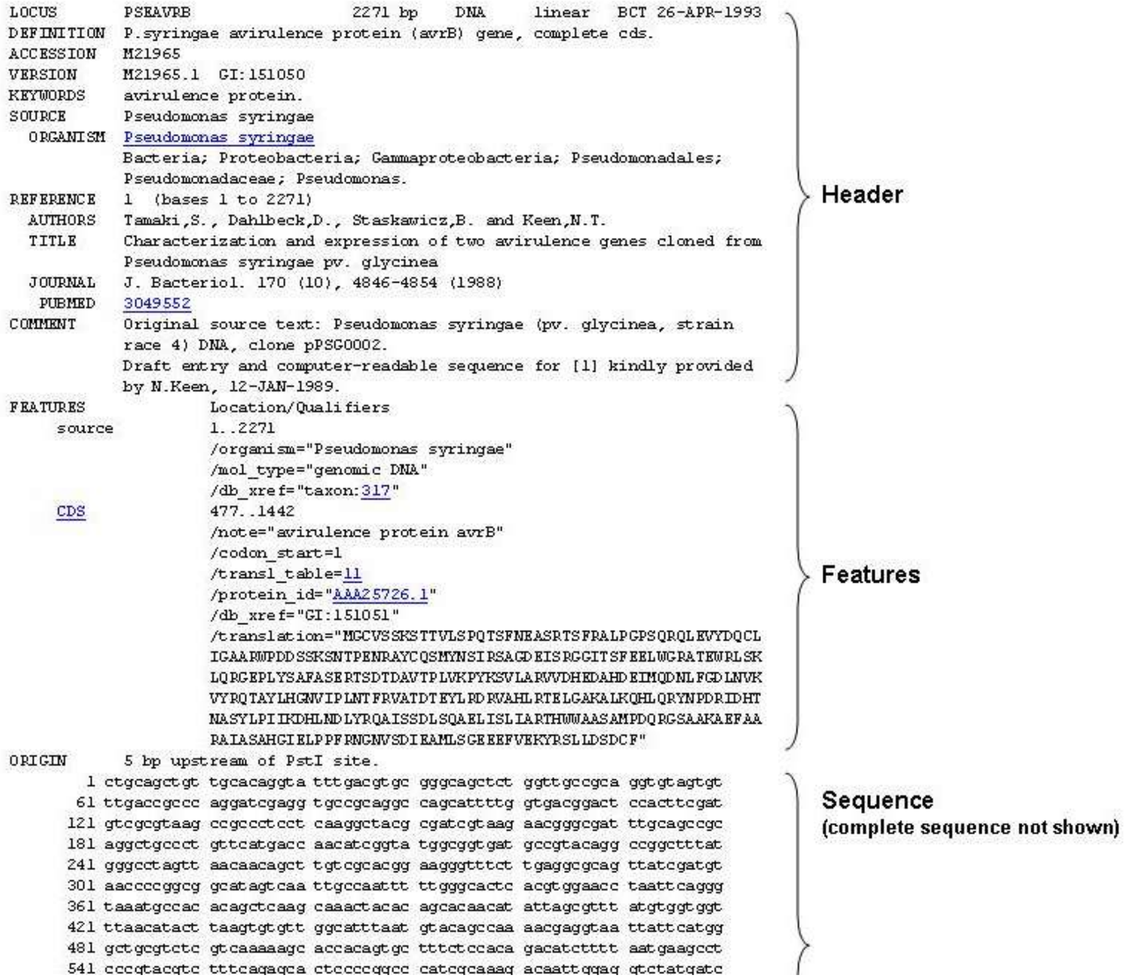
Automatic annotation

from fasta (.fa) to genbank files (.gb)

Automatically **finding** the genetic features possibly encoded on the contigs through alignment with a database of orthologous genes

Annotating the predicted genetic features

In series **for each contig** (separated by //)



```
LOCUS       PSEAVRB                 2271 bp    DNA     linear   BCT 26-APR-1993
DEFINITION  P.syringae avirulence protein (avrB) gene, complete cds.
ACCESSION   M21965
VERSION     M21965.1  GI:151050
KEYWORDS    avirulence protein.
SOURCE      Pseudomonas syringae
  ORGANISM  Pseudomonas syringae
            Bacteria; Proteobacteria; Gammaproteobacteria; Pseudomonadales;
            Pseudomonadaceae; Pseudomonas.
REFERENCE   1  (bases 1 to 2271)
  AUTHORS   Tamaki,S., Dahlbeck,D., Staskawicz,B. and Keen,N.T.
  TITLE     Characterization and expression of two avirulence genes cloned from
            Pseudomonas syringae pv. glycinea
  JOURNAL   J. Bacteriol. 170 (10), 4846-4854 (1988)
   PUBMED   3049552
COMMENT     Original source text: Pseudomonas syringae (pv. glycinea, strain
            race 4) DNA, clone pPSG0002.
            Draft entry and computer-readable sequence for [1] kindly provided
            by N.Keen, 12-JAN-1989.
FEATURES             Location/Qualifiers
     source          1..2271
                     /organism="Pseudomonas syringae"
                     /mol_type="genomic DNA"
                     /db_xref="taxon:317"
     CDS             477..1442
                     /note="avirulence protein avrB"
                     /codon_start=1
                     /transl_table=11
                     /protein_id="AAA25726.1"
                     /db_xref="GI:151051"
                     /translation="MGCVSSKSTTVLSPQTSFNEASRTSFRALPGPSQRQLEVYDQCL
                     IGAARWPDDSSKSNTPENRAYCQSMYNSIRSAGDEISRGGITSFEELWGRATEWRLSK
                     LQRGEPLYSAFASERTSDTDAVTPLVKPYKSVLARVVDHEDAHDEIMQDNLFGDLNVK
                     VYRQTAYLHGNVIPLNTFRVATDTEYLRDRVAHLRTELGAKALKQHLQRYNPDRIDHT
                     NASYLPIIKDHLNDLYRQAISSDLSQAELISLIARTHWWAASAMPDQRGSAAKAEFAA
                     RAIASAHGIELPPFRNGNVSDIEAMLSGEEEFVEKYRSLLDSDCF"
ORIGIN      5 bp upstream of PstI site.
        1 ctgcagctgt tgcacaggta tttgacgtgc gggcagctct ggttgccgca ggtgtagtgt
       61 ttgaccgccc aggatcgagg tgccgcaggc cagcattttg gtgacggact ccacttcgat
      121 gtcgcgtaag ccgccctcct caaggctacg cgatcgtaag aacgggcgat ttgcagccgc
      181 aggctgccct gttcatgacc aacatcggta tggcggtgat gccgtacagg ccggctttat
      241 gggcctagtt aacaacagct tgtcgcacgg aagggtttct tgaggcgcag ttatcgatgt
      301 aaccccggcg gcatagtcaa ttgccaattt ttgggcactc acgtggaacc taattcaggg
      361 taaatgccac acagctcaag caaactacac agcacaacat attagcgttt atgtggtggt
      421 ttaacatact taagtgtgtt ggcatttaat gtacagccaa aacgaggtaa ttattcatgg
      481 gctgcgtctc gtcaaaaagc accacagtgc tttctccaca gacatctttt aatgaagcct
      541 cccgtacgtc tttcagagca ctccccggcc catcgcaaag acaattggag gtctatgatc
```

Automatic annotation

Web-server based tools

NCBI Prokaryotic Genomes Automatic Annotation Pipeline service via email	some days
RAST	1 day

Local tools

PROKKA	10 minutes

Prokka is a command line based tool; possible to install it on Galaxy; running on ARIES

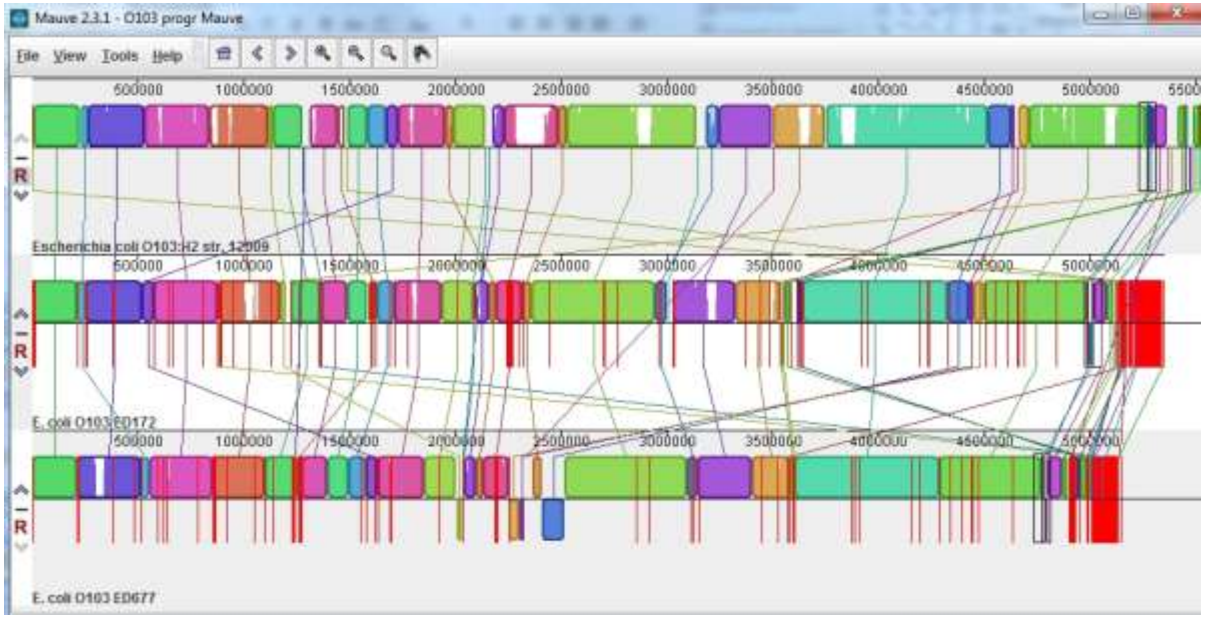
Multiple genome comparison

Possibility to **visually inspect** the obtained contigs for the presence of interesting **genomic regions** such as bacteriophages and pathogenicity islands (LEE, OI-122 and OI-57)

Ref

Contigs strain 1

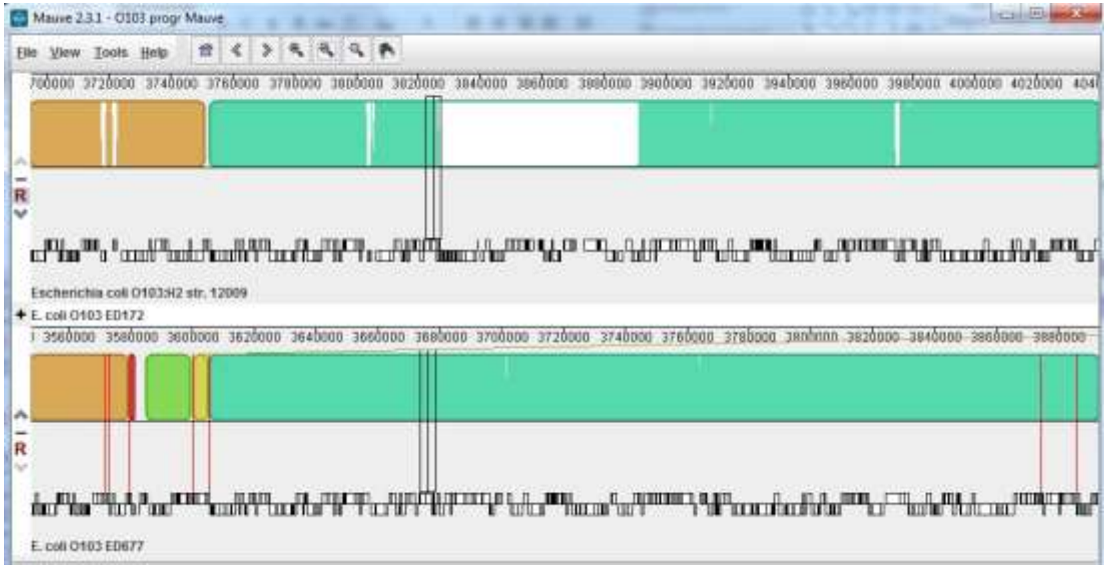
Contigs strain 2



Perform **multiple progressive alignments** of the draft genomes of several test strains (as ordered contigs) on the reference sequence

Multiple genome comparison

When using an **annotated reference sequence (.gb file)**, possibility to zoom in to inspect the genes



White regions: absent in the contigs in analysis

MAUVE is currently a local running tool, but it is in implementation for galaxy; easy to use graphical interface