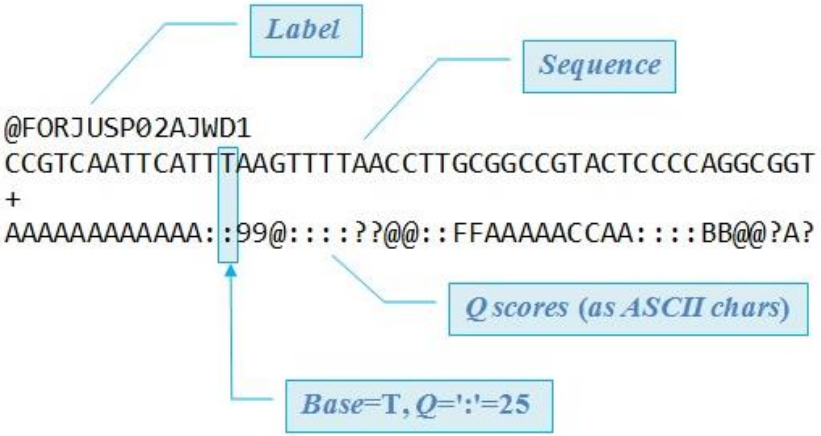
Basic tools for NGS data mining: quality check, assembly, annotation, alignment and blast

Valeria Michelacci

NGS course, June 2016

Quality check

Output of NGS sequencers



.fastq file

Input for quality check

Sequencing errors would impact every following application

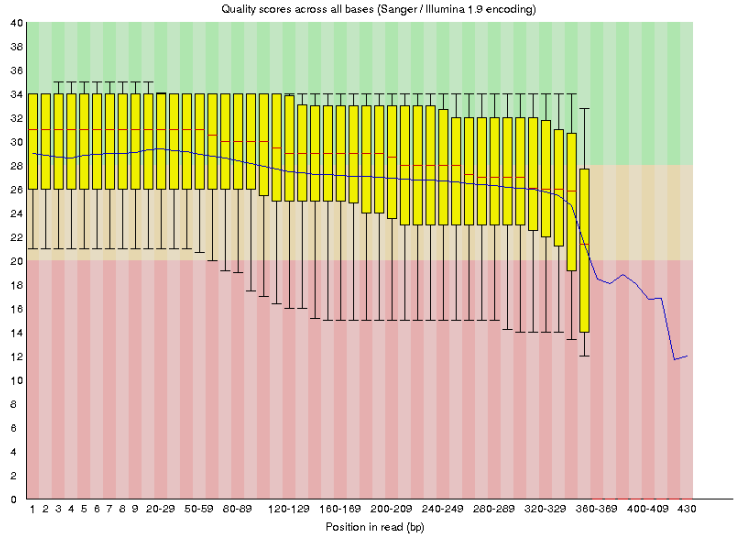
Unreliability of following results (and difficulty to detect the existence of problems!)

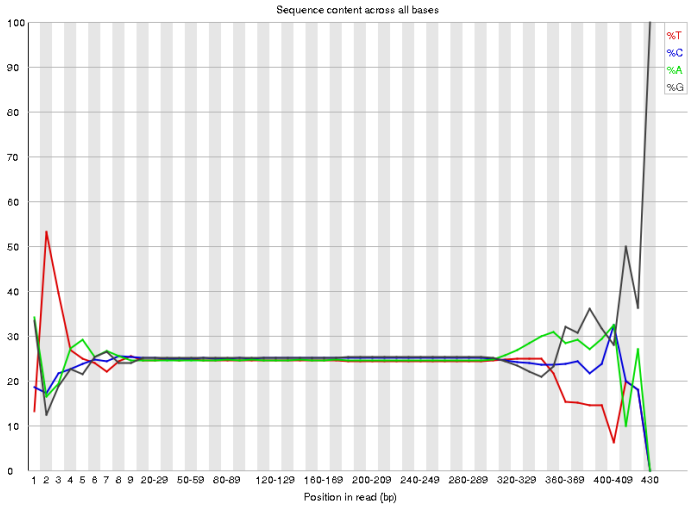
Parameters to control

- Phred score
- GC content distribution over all sequences
- Distribution of undetermined bases (N)
- Sequence Duplication Levels
- ★ Length of the reads
- ★ Coverage

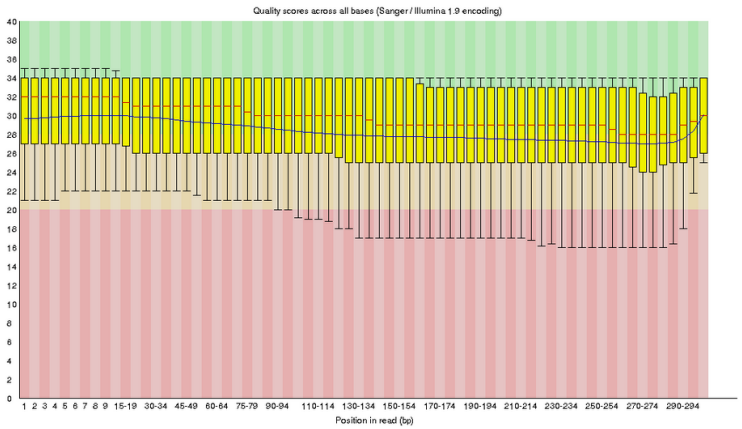
Adoption of corrective actions is possible to minimize some of these problems

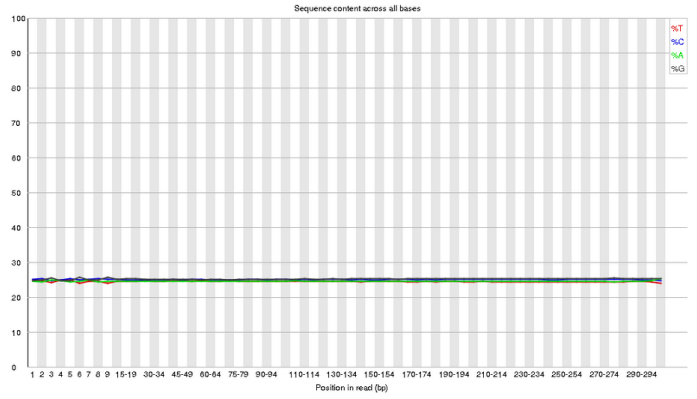
Tool for quality check





Positional and quality trimming





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

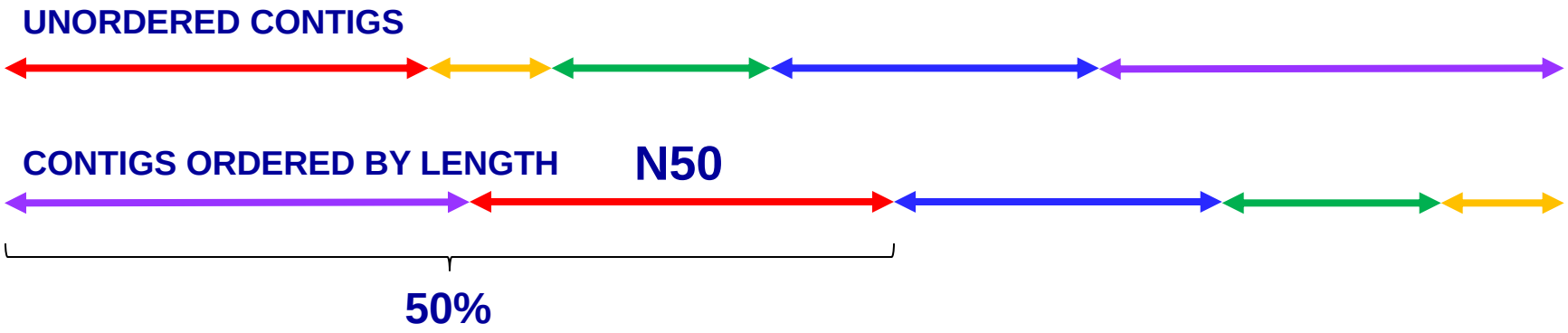
```
>NODE_1_length_449_cov_4.835189
ATCTTTCGCGCCTTCCAGCTCCAGCCATTCGGAACCGTTCGCCAGAAAACGGGCGTAATC
GGGTAAGACATAGCGCGGTTTGTACGGCGCATGACCTTCAAACATATCGCAGATTACACC
TTCATCCAGCGCGCGGCGGGCTTCGGCAGGAAGCTGTGGGTAAGGCAGATTGTTTTCTGC
TTCCAGTGCCAGAAAATGGCGCTTCTGCTCCGGGCTAAGCACTGGGCTGGTGACAATTTG
CTGGCAACGTTGTTGCAGTGCATTTTCATGAGAAGTGGGCATCTTCTTTTCCTTTTATGC
CGAAGGTGATGCGCCATTGTAAGAAGTTTCGTGATGTTCACTTTGATCCTGATGCGTTTG
CCACCACTGACGCATTCATTTGAAAGTGAATTATTTGAACCAGATCGCATTACAGTGATG
CAAACTTGTAAGTAGATTTCCTTAATTGTGATGTGTATCGAAGTGTGTTGCGG
>NODE_2_length_309_cov_4.686084
ACTGGTCAGTGCGGGTATCCTTGGACAATGGCCGATTGGACGTCTGGCGGATAAGTTTGG
TCGACTGCTGGTGTTGCGTGTTCAGGTCTTTGTCGTCATTCTCGGCAGTATCGCGATGCT
TAGCCAGGCGGCGATGGCCCCAGCGTTATTCATCCTCGGTGCCGCTGGCTTTACGCTATA
TCCGGTGGCGATGGCATGGGCTTGCGAGAAAGTTGAACATCATCAACTGGTGGCGATGAA
CCAGGCCTTACTGTTGAGCTATACTGTGGGAAGTCTGCTTGGCCCGTCATTTACCGCTAT
GCTAATGCAGAATTTCTCCGATAATTTATTGTT
>NODE_3_length_101_cov_3.346535
AGCGCATGAGCGCGCAGCGCCGCCGTTACGTGGTGCATCAGCATGATGTTGGCCGGAGAG
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

~10 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:

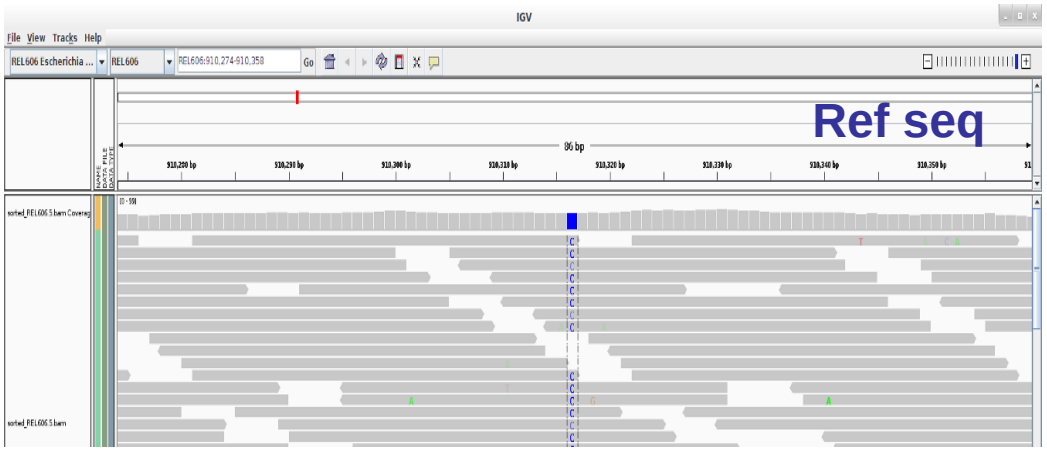
Maximum contig lenght

Coverage of the contigs

Consensus lenght

Alignment (mapping)

Alignment of the sequencing reads on a reference sequence or on a database of reference sequences



Possibility to directly inspect the presence/absence of a target sequence and the presence of SNPs at interesting positions by opening the bam file with a graphic viewer (e.g. IGV)

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

Possibility to convert the output in a sam file (tabular) to extract interesting info and sequences

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)	

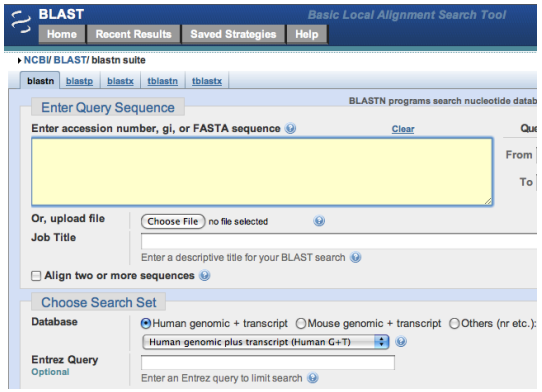
What we need now:

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

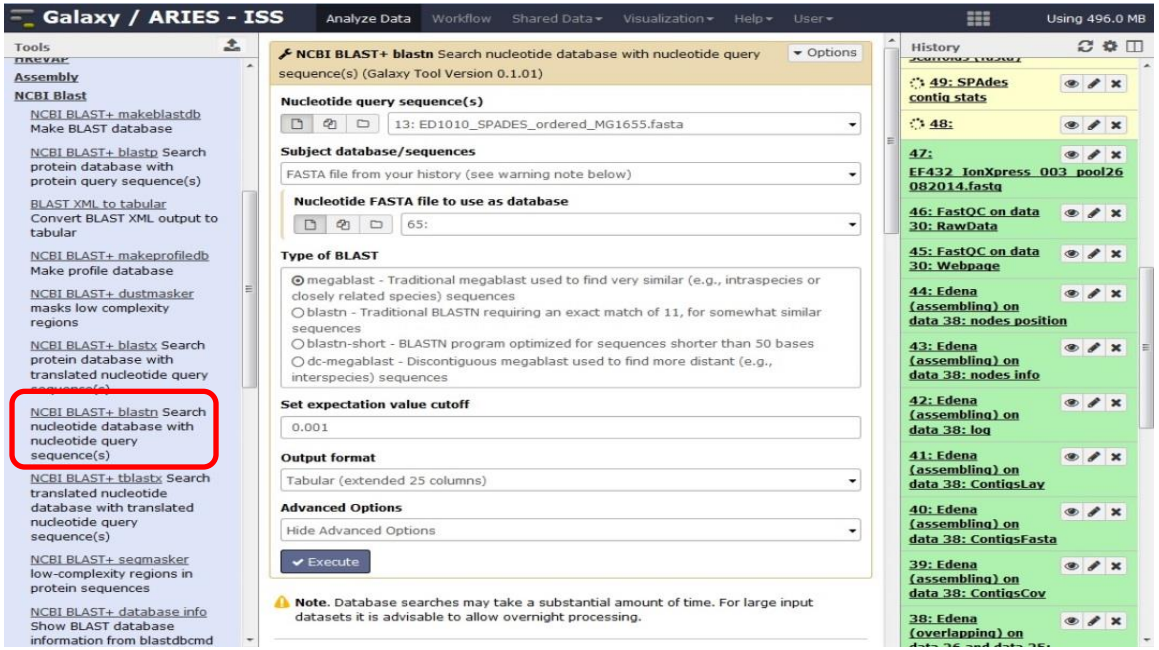
```
blastn -query text_query.txt -db refseq_rna.00 -out output.txt
```

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
atgaaaatatctcttgtcgttcctgtcttcaatgaagaagaagcgataccctgttttctat
aaaacggtacgtgaattccaagagttgaagccatatgaagtagaaattgtattcataaaat
gacggaagtaaagatgccacagagtcaattattaacgcgctggctgtttcagacccgcta
gttgttccgctgtcatttacacgcaacttttggtaaagaaccagccttatttgcagggtta
gaccatgcaagcggcgatgctgtaattcctattgatgtcgacctgcaagacccaattgag
gttatccctcatcttattgaaaagtggcaggcaggtgctgacatggtgcttgctaaacgt
tcagaccgctcaactgatggacgactgaaacgtaagacagctgagtggttctataaatta
cacaacaaaataagcaccccaaagatcgaggaaaatgtcggagattttcgactcatgtct
cgtgaggttgtggagaacattaaactgttgcctgagcgcaatcttttcatgaaaggcata
ctgagctgggtgggtggtcagacggatgtcgttgaatatgtacgcgcagagcgtgttgct
ggcatctcaaaatttaatggctggaaattatggaatctggcactggaaggtatcacaagt
ttttcaacctttcctcttcgcgtatggacttatataggcttgtttgttgcaagcatttca
tttttatatggtgcatggatgattatagacacccttgtctttggtaacccagtacgcgggg
tatccctccctgcttgtatcaatacttttcttgggtggagtgcaactgatcgggattggt
gttctcggagaatatataggtagaatctatattgaagtaaaaaatagacccaaatacatc
ataaaaaaatctcatcgaggtaacccatga
>gtrII_24111749
atgattaaaattaatttattaaaaaatgcaaatctccttgccttcatatcatgttttgca
atatcaatttattgttattgggggtggttgtacgatggaacacttaatattgatggcgag
tttacaaataatttttatcaaacaatcacgcttgggcggtggtttcacacttttttgcga
cattactttctccctgagccttttcactttatataacaccattaatagccttatctttt
atcatcatttcagcatttataatctgcagatcgctaaagctagaatcttatgaattattg
ataggtatgttagtatttattaccttccctcagatctcctatcaattagagtttcttaac
caagctgatactgtgggaattgcttttctactggcagcgatatcagcaattatttttcac
tcgcaaaaaaaataggattgtgatatttctggtatagtactgtccattctttcaatggca
atctaccaaacattcgtaacatatattattgcattcgtcattgggttgcagataaattcg
ataatacgaaatgagaaaaatattcgtgaatctttttatagttcatgtttatctctatcc
ctcatagctttatctaccttaatttacctgctattaaccaaagctatcaagcattatttt
tcgcttgaatcgaacgagtacatctcaaattatatacaaaatgcaagcgatattaaatgg
cttgttaaatcagccatagataatatatataacttctataacaatcctcccactggttta
aacctatacaagtggttactgattcctttattaattctgatgtttaccctaacatataaa
ttaaaaacaagatcaatttatttgatttcatcaatcattttcatttatatacttccggtt
atatttatcgttgttgttggctcaggggcgccacctcgcctgtttgttttaatgcctata
gtagcagtaattttgttttcttgcttaagcaattttcgctctataaaatacctaaactgc
atgttttttttatttattatatatttaatggcgtttcaacatccaaaaatctatttttgaat
gatactctcgcaagacagaaagatatctctttagctaaagaaatatcatacacatctcaa
acaaaaggcatttcccttaacggaaaatatatatatatatatatggttcaaacgactcagga
aatatgctctccatgagcgcagacactttttggaaaatctttttttttggtgggatggtggc
aactattttaggatggttgcatttatgaattactatggaatctgtaattgcaaaccagca
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