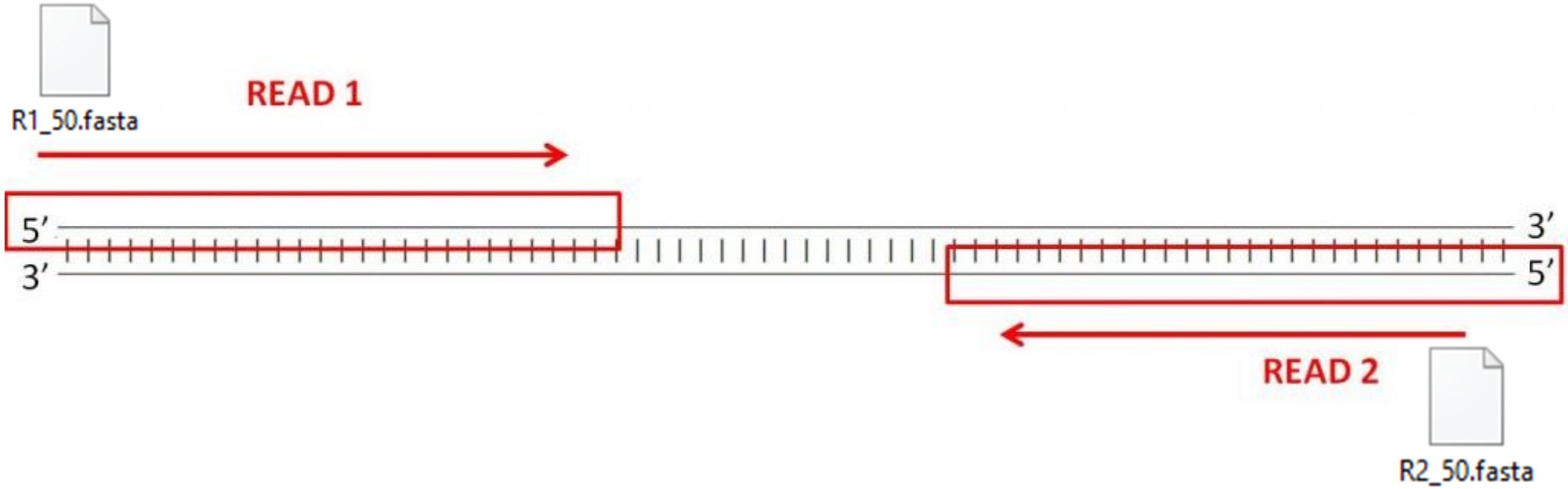
Genome assembly and annotation

The presente workflow was processed in Linux ubunto.

Introduction

- This work consists of assembling a genome using two reads, a forward and a reverse in order to identify the organism and the number of genes present in that genome.



Quality control- Fastx_toolkit

FASTX-Toolkit – Performs pre-processing tasks before mapping strings to produce better mapping results.

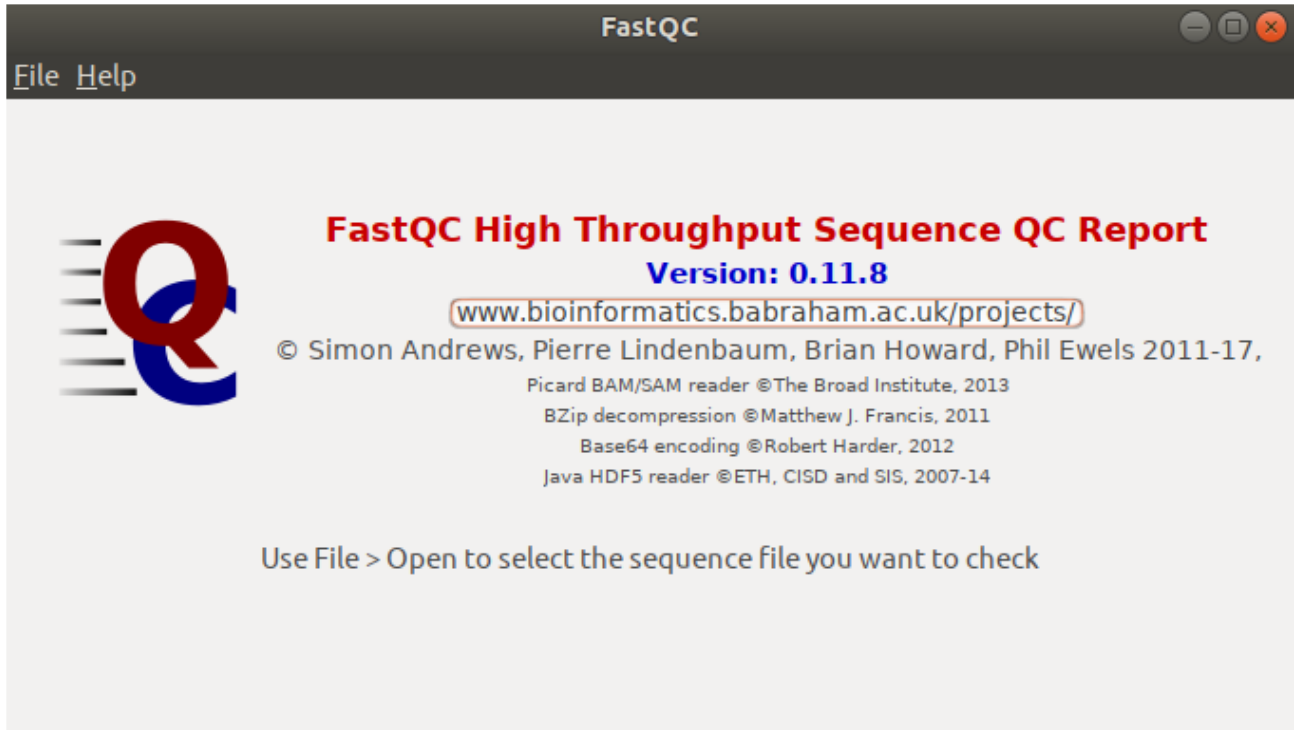
The toolkit's fastx_trimmer command was used to make adjustments to the two reads

```
> fastx_trimmmer –i R1_50.fasta –o R1_trim.fasta
> fastx_trimmmer –i R2_50.fasta –o R2_trim.fasta
```

Trimming will improve the quality and speed of analysis, removing unnecessary parts of the two reads.

Quality control- FastQC

Next, FastQC was used to confirm the quality of the two reads



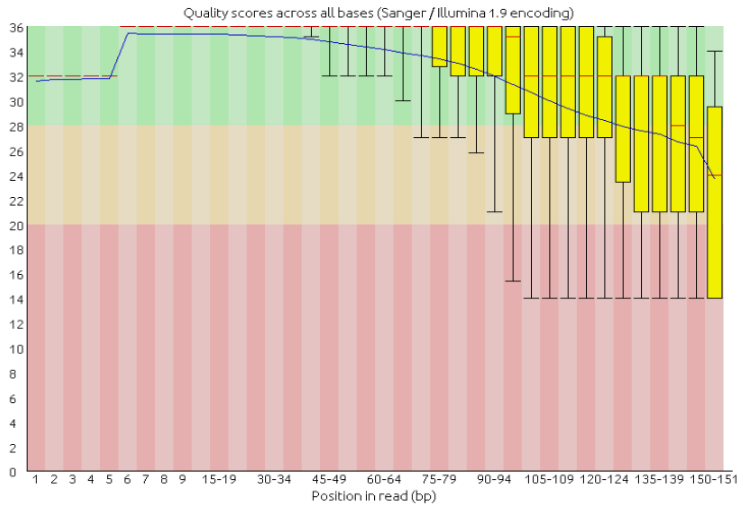
FastQC – performs quality control checks on raw sequence data from high-throughput sequences.

Quality control- FastQC

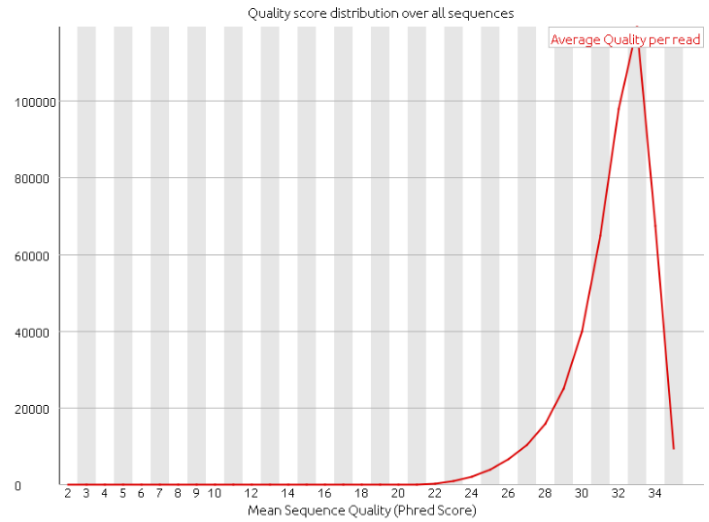
R1

Measure	Value
Filename	R1TRIM.fasta
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	465661
Sequences flagged as poor quality	0
Sequence length	35-151
%GC	36

Per base sequence quality



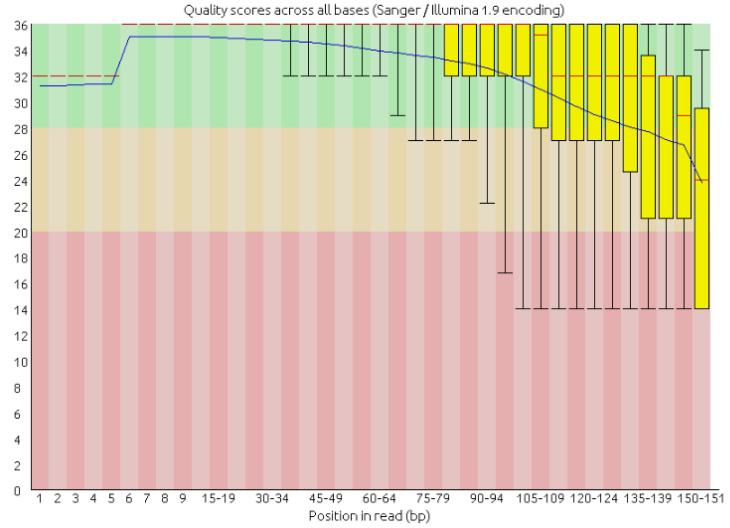
Per sequence quality scores



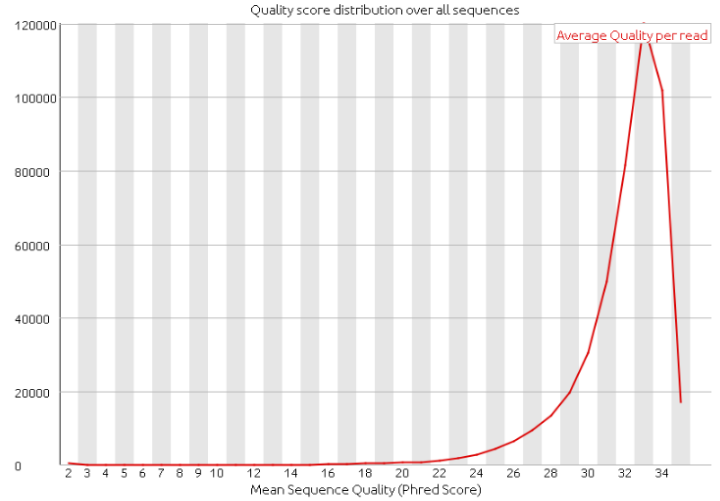
R2

Measure	Value
Filename	R2TRIM.fasta
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	465661
Sequences flagged as poor quality	0
Sequence length	35-151
%GC	37

Per base sequence quality



Per sequence quality scores

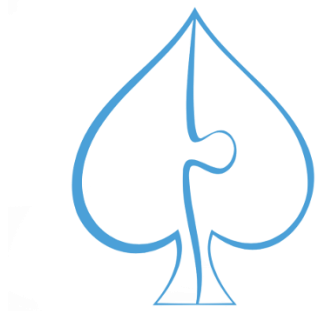


Quality control allows you to check whether there is a need for data improvement or not. From the graphs it is possible to observe that the quality of the data is good so that they can be followed for analysis as they present values around 33.

Assembly - SPAdes

SPAdes – allows to assemble the genomic sequences.

It is capable of producing long, highly accurate contigs but is not suitable for use with large and complex genomes.



Comand to execute:

> spades.py –t 2 –k 21,33,55,77,99,127 –m 2 --only-assembler --pe1-1 R1TRIM.fasta --pe1-2 R2TRIM.fasta –o spadesassembly

Assembly - MEGAHIT

Megahit – allows to assemble the genomic sequences.

Suitable for large and complex genomes and works faster than SPAdes.

Comand to execute:

> megahit –t 2 –m 2 --k-list 21,33,55,77,99,127 -1 R1TRIM.fasta -2 R2TRIM.fasta –o HITassembly

Avaliation - MetaQUAST

- **MetaQUAST** – evaluates and compares metagenome assemblies based on contig alignments of a reference

Comand to execute:

```
> quast contigs.fasta --gene-finding
```

- The -'gene-finding 'command will use the GeneMark tool to predict the number of genes present in the genome

SPAdes

Statistics without reference	contigs
# contigs	182
# contigs (>= 0 bp)	227
# contigs (>= 1000 bp)	159
# contigs (>= 5000 bp)	121
# contigs (>= 10000 bp)	96
# contigs (>= 25000 bp)	56
# contigs (>= 50000 bp)	30
Largest contig	163 633
Total length	4 314 728
Total length (>= 0 bp)	4 326 679
Total length (>= 1000 bp)	4 297 905
Total length (>= 5000 bp)	4 214 527
Total length (>= 10000 bp)	4 032 650
Total length (>= 25000 bp)	3 329 357
Total length (>= 50000 bp)	2 392 372
N50	55 076
N75	28 258
L50	26
L75	53
GC (%)	37.24
Mismatches	
# N's	0
# N's per 100 kbp	0
Predicted genes	
# predicted genes (unique)	3916
# predicted genes (>= 0 bp)	3876 + 41 part
# predicted genes (>= 300 bp)	3360 + 26 part
# predicted genes (>= 1500 bp)	565 + 5 part
# predicted genes (>= 3000 bp)	75 + 1 part

MEGAHIT

Statistics without reference	final.contigs
# contigs	183
# contigs (>= 0 bp)	222
# contigs (>= 1000 bp)	152
# contigs (>= 5000 bp)	107
# contigs (>= 10000 bp)	87
# contigs (>= 25000 bp)	53
# contigs (>= 50000 bp)	31
Largest contig	170 723
Total length	4 310 684
Total length (>= 0 bp)	4 323 837
Total length (>= 1000 bp)	4 287 329
Total length (>= 5000 bp)	4 166 361
Total length (>= 10000 bp)	4 017 255
Total length (>= 25000 bp)	3 437 164
Total length (>= 50000 bp)	2 662 350
N50	59 003
N75	31 609
L50	22
L75	46
GC (%)	37.23
Mismatches	
# N's	0
# N's per 100 kbp	0
Predicted genes	
# predicted genes (unique)	3906
# predicted genes (>= 0 bp)	3863 + 51 part
# predicted genes (>= 300 bp)	3336 + 34 part
# predicted genes (>= 1500 bp)	573 + 6 part
# predicted genes (>= 3000 bp)	74 + 3 part

Anotation - Prokka

Prokka - performs quick annotation of a genome, identifying and separating existing proteins.

Comand to execute:

› **prokka contigs.fasta**

Prokka Output ->

```
>IDMIFEGJ_00001 Ribonuclease 3
MKINLERLCRRLNYQFKNTAYLKQALTHCSFGSDNYERFEFLGDSILSFVIANELFHRFP
LQSEGQLSRLRSFLVRGDMLLELAKEIELGDFLYLGQGELKSGGFRRASILSDALEAVFA
AIFLDGGVDSAKEVILKLYRSRLEDPNLNDCLKDAKTQLQEYLQAEKIALPEYKLTKVEG
DEHEQIFYIICTVDGVKKETFGQGSNRRKAEQLAAQAMLKSLRSGD
>IDMIFEGJ_00002 hypothetical protein
MNKNQGMTFIGTLFTIAVVVMVATIIMRVVPVYLQYYAIIESVKGLNTISQSSLTGDSIQ
DVMVLKSDLDKRLDINGVSSLKDNQLTIEPHGPNKFIVKIKYQVTRPLVSNVSLLFDFNH
TEEVVAGSEN
>IDMIFEGJ_00003 Signal peptidase I
MNFALILVILSFISGFIYLLDVLFWAKKRTEDQQPNRIIEYSRSFFPVFFIVLLLRSFLI
EPFRIPSGSLEPTLLVGDFVAVNKFIYGLRLPVWEKKVVSIANPKTGDVSVFRWPPDPSF
DYIKRVIGVPGDKISYHNKVLTVNGKEAKQTFVEYTIDESSGKAVSKYKEDLNGTVHDIF
VRTNVPAVDFDLVVPQGNYFMMGDNRDDSADSRFWGFVPDSYLRGKAFLVWMSWNSNTAN
VRWSKIGKFIP
```

Anotation - Prokka

Prokka also creates a text file that gives additional information about the genome, such as the number of total genes, obtaining a number close to the results of metaQUAST.

SPAdes

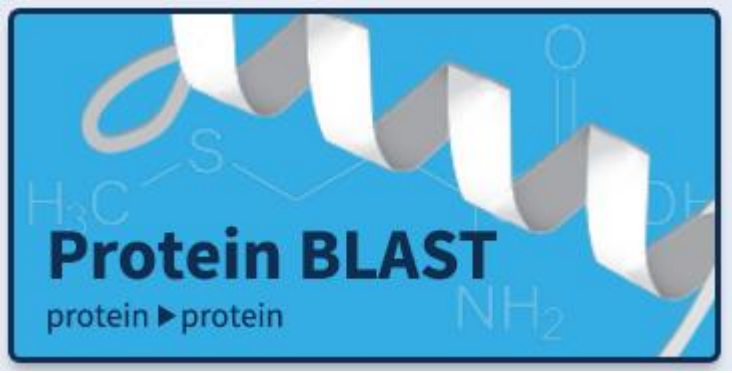
```
organism: Genus species strain
contigs: 227
bases: 4326679
CDS: 3839
rRNA: 6
tRNA: 41
tmRNA: 1
```

MEGAHIT

```
organism: Genus species strain
contigs: 222
bases: 4323837
CDS: 3832
rRNA: 5
tRNA: 43
tmRNA: 1
```

Resultados – BLAST

BLAST - Compare sequences of nucleotides or proteins with sequences stored in databases.



A BLAST was made with the first 10 proteins in the file created by the prokka.

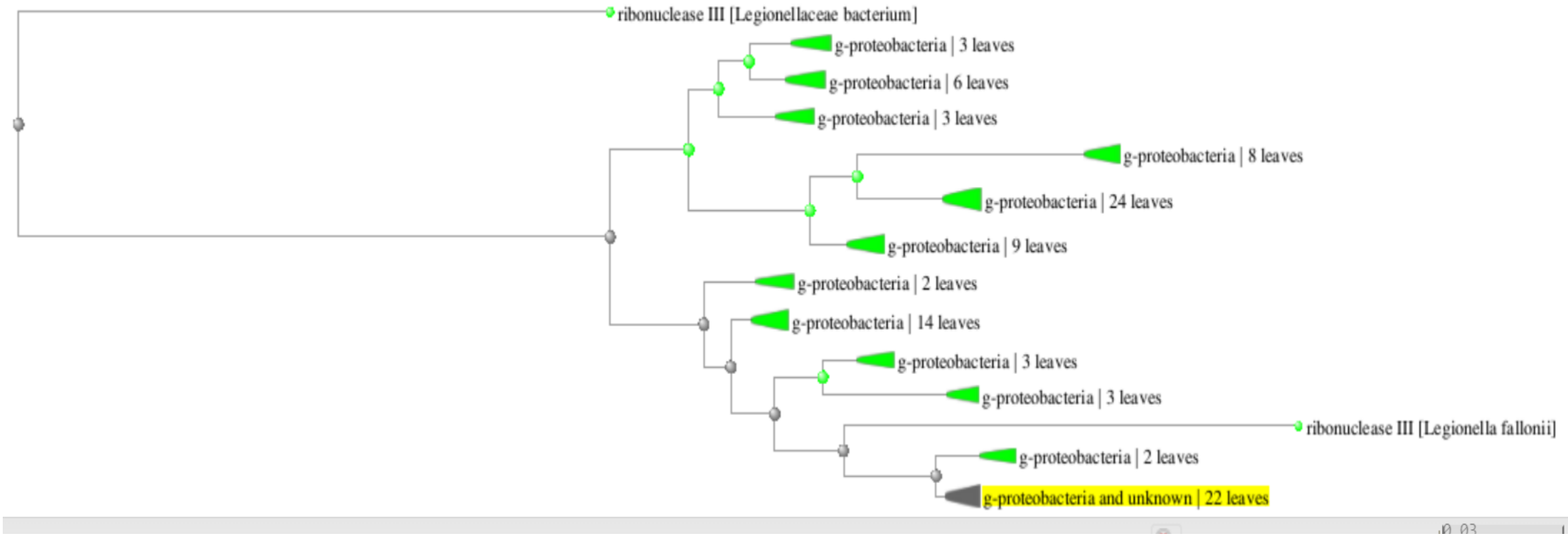
Results – BLAST (SPAdes)

It was observed that the organism is a legionella

Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
ribonuclease III [Legionella steigerwaltii]	412	412	99%	1e-144	87.95%	WP_058476480.1
ribonuclease III [Legionella steelei]	410	410	99%	8e-144	87.11%	WP_058510050.1
ribonuclease III [Legionella santicrucis]	408	408	99%	3e-143	88.44%	WP_058513158.1
ribonuclease III [Legionella sainthelensi]	408	408	100%	5e-143	87.61%	WP_027270213.1
ribonuclease III [Legionella gratiana]	408	408	99%	5e-143	87.95%	WP_058498376.1

Organism	Blast Name	Score	Number of Hits
Legionellales	g-proteobacteria		243
. Legionellaceae	g-proteobacteria		241
. . Legionella	g-proteobacteria		3
. . . Legionella steigerwaltii	g-proteobacteria	412	3
. . . Legionella steelei	g-proteobacteria	410	2
. . . Legionella sp. 39-23	g-proteobacteria	410	1
. . . Legionella santicrucis	g-proteobacteria	408	2

Results – BLAST (SPAdes)



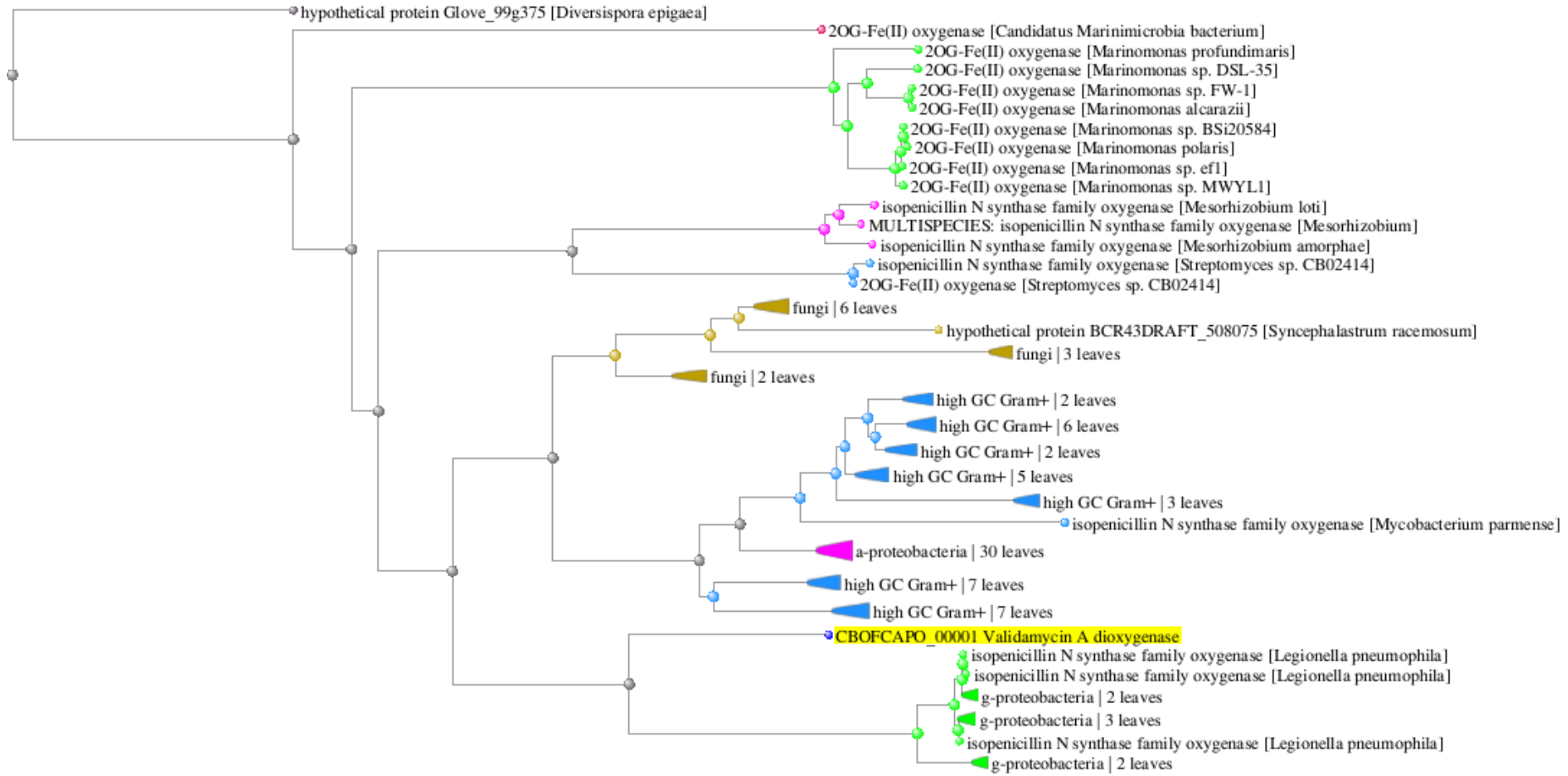
Resultados – BLAST (MEGAHIT)

It was observed also with MEGAHIT that the organism is a legionella.

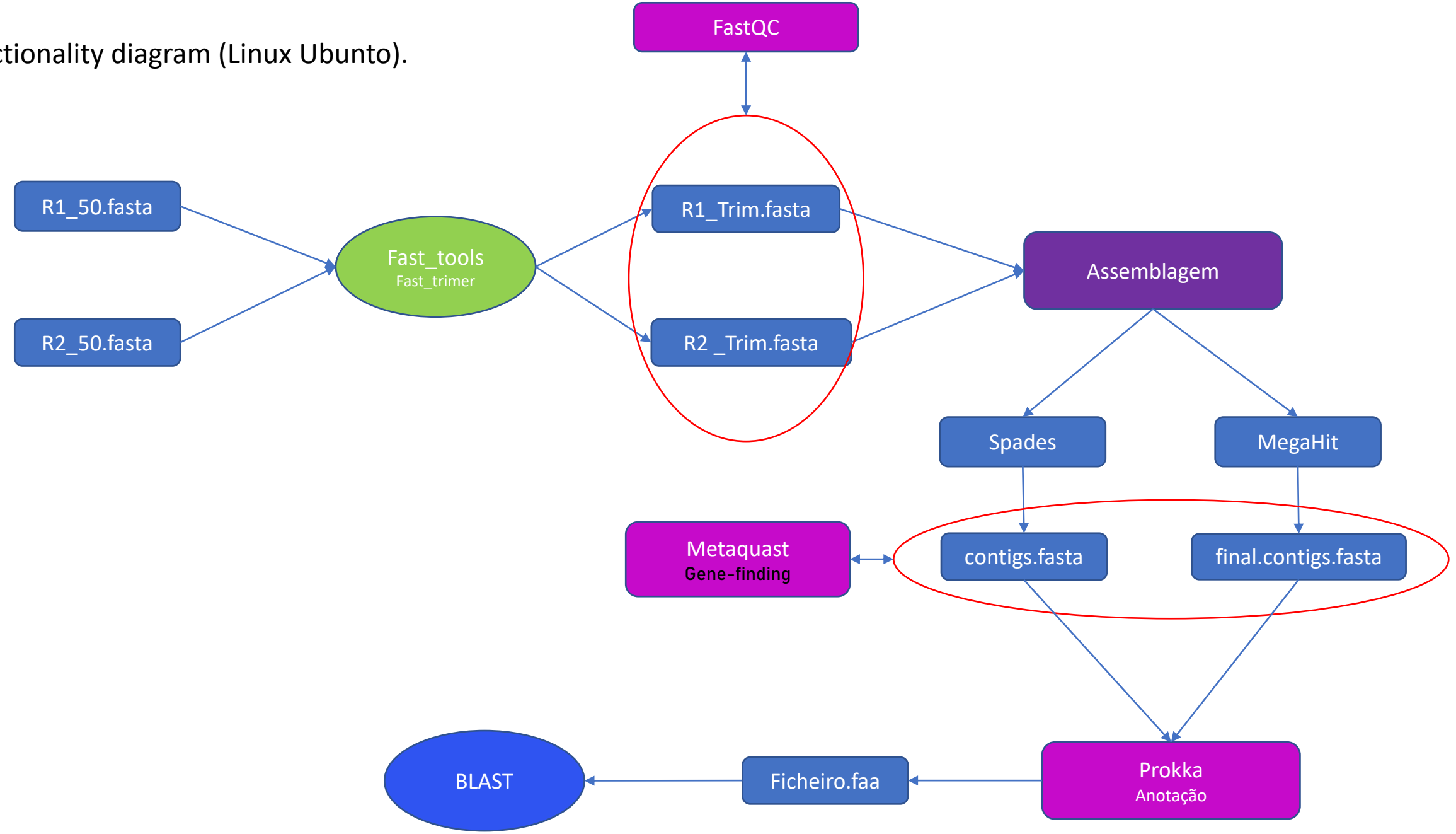
Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
isopenicillin N synthase family oxygenase [Legionella pneumophila]	430	430	99%	7e-148	59.23%	WP_106183721.1
isopenicillin N synthase family oxygenase [Legionella pneumophila]	429	429	99%	1e-147	59.23%	WP_080464050.1
isopenicillin N synthase family oxygenase [Legionella pneumophila]	429	429	99%	2e-147	58.93%	WP_106194543.1
isopenicillin N synthase family oxygenase [Legionella pneumophila]	428	428	99%	2e-147	59.23%	WP_050598185.1
isopenicillin N synthase family oxygenase [Legionella pneumophila]	428	428	99%	3e-147	59.23%	WP_106225228.1
isopenicillin N synthase family oxygenase [Legionella pneumophila]	427	427	99%	1e-146	58.93%	WP_106221572.1

Organism	Blast Name	Score	Number of Hits	Description
cellular organisms			185	
. Bacteria	bacteria		168	
. . Proteobacteria	proteobacteria		101	
. . . Gammaproteobacteria	g-proteobacteria		34	
. . . . Legionella	g-proteobacteria		21	
. . . . . Legionella pneumophila	g-proteobacteria	430	18	Legionella pneumophila hits
. . . . . Legionella pneumophila subsp. fraseri	g-proteobacteria	423	2	Legionella pneumophila subsp. fraseri hits

Resultados – BLAST (MEGAHIT)



Functionality diagram (Linux Ubunto).



References

- Genome trimming - https://doi.org/10.1371/journal.pone.0085024
- Fastx_toolkit - https://anaconda.org/bioconda/fastx_toolkit
- SPAdes - http://cab.spbu.ru/software/spades/
- MEGAHIT - https://kbase.us/applist/apps/MEGAHIT/run_megahit/release
- http://www.metagenomics.wiki/tools/assembly/megahit
- Quast - https://github.com/ablab/quast
- GeneMark - http://exon.gatech.edu/GeneMark/
- Prokka - https://github.com/tseemann/prokka
- BLAST - https://blast.ncbi.nlm.nih.gov/