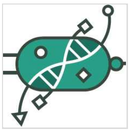


Oct 27, 2022

Bacterial genome annotation script using BLASTN

Ana Mariya Anhel[1], Lorea Alejaldre[1], Ángel Goñi-Moreno[1]

[1]Centro de Biotecnología y Genómica de Plantas, Universidad Politécnica de Madrid (UPM)-Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria (INIA/CSIC), Madrid, Spain

Ángel Goñi-Moreno: angel.goni@upm.es

1 Works for me   Share

dx.doi.org/10.17504/protocols.io.dm6gpjrb1gzp/v1

biocomp.cbgp

ABSTRACT

This protocol uses a python based script and command-line blastn to annotate Sanger sequencing results from genome amplifications. Its main use in our lab (https://biocomputationlab.com) is to identify the location and gene locus of transposon inserts in microbial bacterial genomes of *Pseudomonas putida* KT2440. However, this script can be used for other bacterial genomes for which its genome sequence and annotation are available.

Script was developed in python 3.9 with blastn version 2.2.18.

DOI

dx.doi.org/10.17504/protocols.io.dm6gpjrb1gzp/v1

PROTOCOL CITATION

Ana Mariya Anhel, Lorea Alejaldre, Ángel Goñi-Moreno 2022. Bacterial genome annotation script using BLASTN. **protocols.io**
https://dx.doi.org/10.17504/protocols.io.dm6gpjrb1gzp/v1

FUNDERS ACKNOWLEDGEMENT

Comunidad de Madrid
Grant ID: Y2020/TCS-6555, 2019-T1/BIO-14053

MCIN/AEI
Grant ID: CEX2020-000999-S, PID2020-117205GA-I00

European Research Council
Grant ID: 101044360

KEYWORDS

Genome anotation, Bacterial, P. putida, Transposon, Transposon library, E. coli

CREATED

Sep 22, 2022

LAST MODIFIED

Oct 27, 2022

OWNERSHIP HISTORY

Sep 22, 2022   Lorea Alejaldre

Oct 27, 2022   biocomp.cbgp

PROTOCOL INTEGER ID

70385

PARENT PROTOCOLS

In steps of

High-throughput workflow for the genotypic characterization of transposon library variants

GUIDELINES

This script needs 4 arguments in the following order:

1. Directory of folder containing sequencing reads in .txt or .seq format
2. Reads file type (txt or seq)
3. Genome file to perform blastn alignment (FASTA format)
4. Genome annotation file (.csv)

MATERIALS TEXT

Software

- python 3.9.10
- python packages: sys, pandas, os and subprocess
- blastn 2.10.0+ (https://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/)

BEFORE STARTING

To run this script command-line blastn and python 3 with packages sys, pandas and os must be installed.

Annotation of sequencing reads

1 Download genome file in FASTA format and annotation file in .csv for the microbial organism to use as reference

> *Pseudomonas* genome and annotation files can be found in https://www.pseudomonas.com.

2 Run the following python based script with the required arguments

> Command to run blastn annotation script
>
> **python alignment_and_annotation_blastn.py [directory of sequencing reads] [type of file] [genome file in fasta format] [annotation file in csv format]**

> Updated versions of this script can be found in Biocomp GitHub folder

3 Output is a folder named *results_script_blast* which contains three files:

- all_seq_aligned.sam
- all_seq_aligned.txt
- table_reads_genes_description.csv

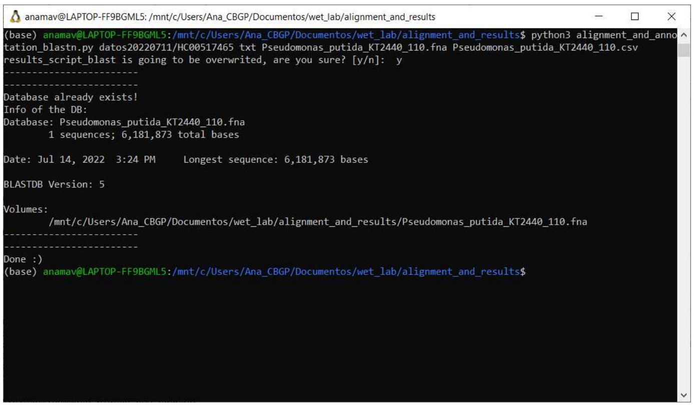
Example: Annotation of sequencing results from *P. putida* KT2440

4 Input files

1. Directory of sequencing reads (it is a zip but shoul be a directory) **HC00517465.zip** In this case the type of file (extension) is txt
2. Genome in FASTA format **Pseudomonas_putida_KT2440_110.fna**
3. Annotation file of that genome **Pseudomonas_putida_KT2440_110.csv**

5 Command-line

Citation: Ana Mariya Anhel, Lorea Alejaldre, ÃÂngel GoÃÂ±i-Moreno Bacterial genome annotation script using BLASTN https://dx.doi.org/10.17504/protocols.io.dm6gpjrb1gzp/v1



anamav@LAPTOP-FF9BGML5: /mnt/c/Users/Ana_CBGP/Documentos/wet_lab/alignment_and_results

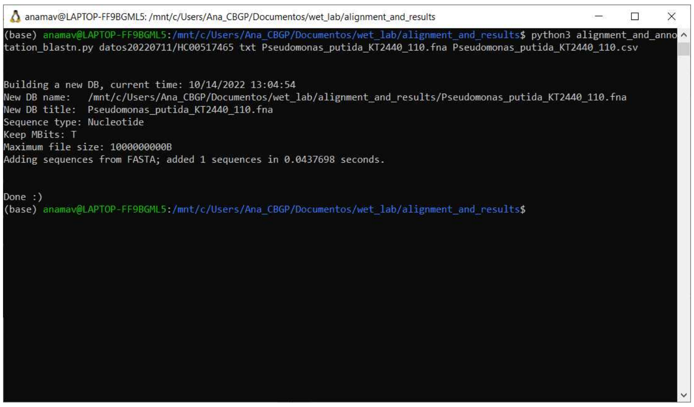
```
(base) anamav@LAPTOP-FF9BGML5:/mnt/c/Users/Ana_CBGP/Documentos/wet_lab/alignment_and_results$ python3 alignment_and_anno
tation_blastn.py datos20220711/HC00517465 txt Pseudomonas_putida_KT2440_110.fna Pseudomonas_putida_KT2440_110.csv
results_script_blast is going to be overwrited, are you sure? [y/n]:  y
------------------------
------------------------
Database already exists!
Info of the DB:
Database: Pseudomonas_putida_KT2440_110.fna
        1 sequences; 6,181,873 total bases

Date: Jul 14, 2022  3:24 PM     Longest sequence: 6,181,873 bases

BLASTDB Version: 5

Volumes:
        /mnt/c/Users/Ana_CBGP/Documentos/wet_lab/alignment_and_results/Pseudomonas_putida_KT2440_110.fna
------------------------
------------------------
Done :)
(base) anamav@LAPTOP-FF9BGML5:/mnt/c/Users/Ana_CBGP/Documentos/wet_lab/alignment_and_results$
```

bash window where the command is executed (the DB was already created and there was an output directory existed also)



anamav@LAPTOP-FF9BGML5: /mnt/c/Users/Ana_CBGP/Documentos/wet_lab/alignment_and_results

```
(base) anamav@LAPTOP-FF9BGML5:/mnt/c/Users/Ana_CBGP/Documentos/wet_lab/alignment_and_results$ python3 alignment_and_anno
tation_blastn.py datos20220711/HC00517465 txt Pseudomonas_putida_KT2440_110.fna Pseudomonas_putida_KT2440_110.csv


Building a new DB, current time: 10/14/2022 13:04:54
New DB name:   /mnt/c/Users/Ana_CBGP/Documentos/wet_lab/alignment_and_results/Pseudomonas_putida_KT2440_110.fna
New DB title:  Pseudomonas_putida_KT2440_110.fna
Sequence type: Nucleotide
Keep MBits: T
Maximum file size: 1000000000B
Adding sequences from FASTA; added 1 sequences in 0.0437698 seconds.


Done :)
(base) anamav@LAPTOP-FF9BGML5:/mnt/c/Users/Ana_CBGP/Documentos/wet_lab/alignment_and_results$
```

bash window where the command is executed without a previously DB created

6 Output files

A new folder named results_script_blast (output files attached in the following zip file) contains a table with information about the alignment and genomic context of each sequencing read.

results_script_blast.zip

Citation: Ana Mariya Anhel, Lorea Alejaldre, ÃÂngel GoÃÂ±i-Moreno Bacterial genome annotation script using BLASTN https://dx.doi.org/10.17504/protocols.io.dm6gpjrb1gzp/v1

query acc.	s. start	% identity	alignment length	mismatches	gap opens	evalue	bit score	subject strand	Locus Tag	Feature Type	Start	End	St
H220707-054_B23_219DZAA034_premix.ab1	6170239	99.04299999999999	209	1	1	1.8599999999999997e-103	374	plus	PP_5408	CDS	61691130	61702940	-
H220707-054_P21_219DZAA035_premix.ab1	6170230	98.618	217	1	2	2.9799999999999995e-106	383	plus	PP_5408	CDS	61691130	61702940	-
H220707-054_L21_219DZAA036_premix.ab1	6170230	99.539	217	0	1	1.4899999999999998e-109	394	plus	PP_5408	CDS	61691130	61702940	-
H220707-054_F19_219DZAA037_premix.ab1	6170230	99.083	218	1	1	1.9699999999999993e-108	390	plus	PP_5408	CDS	61691130	61702940	-
H220707-054_H21_219DZAA038_premix.ab1	6170546	97.22200000000001	108	1	2	1.3e-45	182	minus	PP_5409	CDS	61704660	61723010	-
H220707-054_P19_219DZAA039_premix.ab1	6170546	98.148	108	0	2	2.8e-47	187	minus	PP_5409	CDS	61704660	61723010	-
H220707-054_L19_219DZAA040_premix.ab1	6170547	96.33	109	1	3	6.389999999999999e-44	176	minus	PP_5409	CDS	61704660	61723010	-
H220707-054_N21_219DZAA041_premix.ab1	6170546	99.074	108	0	1	5.679999999999999e-49	193	minus	PP_5409	CDS	61704660	61723010	-
H220707-054_J19_219DZAA046_premix.ab1	6170533	96.84200000000001	95	0	3	9.219999999999999e-38	156	minus	PP_5409	CDS	61704660	61723010	-
H220707-054_D21_219DZAA047_premix.ab1	6170239	99.51700000000001	207	1	0	1.4899999999999996e-104	377	plus	PP_5408	CDS	61691130	61702940	-

Final table of the alignment with the correspondant gene or locus insertion

Citation: Ana Mariya Anhel, Lorea Alejaldre, ÃÂngel GoÃÂ±i-Moreno Bacterial genome annotation script using BLASTN https://dx.doi.org/10.17504/protocols.io.dm6gpjrb1gzp/v1