

Jan Izquierdo

Assignment 1: Identify the origin organism of a fastq sequence

8: Identify the reads origin. (How would you find out from which genome come these reads? To which species they belong? Please describe the method used and the reliability of the results)

```
from Bio import SeqIO
def ex8():
    fastq_file="path/1S-unknown_illumina_2024.fastq" #set path to fastq file
    f=open("path/OUTPUT.fasta", "w")#create new fasta file, with writing permission
    for read in SeqIO.parse(fastq_file, "fastq"):#read the fastq file, and for every read
        print(">" +str(read.id), file=f)#write > and the id
        print(read.seq, file=f)#write the sequence
    f.close()

print(ex8())

#Also can be done using bash command:
#cat 1S-unknown_illumina_2024.fastq | awk '{if(NR%4==1) {printf(">%s\n",substr($0,2));} else if(NR%4==2) print;}' > OUTPUT.fasta

#Create database
#$ makeblastdb -in ref_seq.fasta -dbtype nucl -out reference_database
#Use it
#$ blastn -query OUTPUT.fasta -db reference_database -out results.txt
```

I transformed the fastq into a fasta, then used BLAST to identify a few sequences of the file chosen at random, I searched for the genome of the most common species, the one that appeared most and had lower E values, first it seemed that it was *Limanda limanda*, but with some more samples I ended up choosing *Podarcis lilfordi* because the results had a lower E value. I downloaded the genome of this species, then I transformed the downloaded genome into a BLAST database using the command *makeblastdb* and queried for our sequences in this database using BLAST. In the BLAST search output I could see that most of the searches coincided and that their E values were low, which lets us know that *Podarcis lilfordi* is very likely to be the genome these reads come from.

The methods used were:

- A python program to transform fastq into fasta

- BLAST to identify and corroborate the origin of the sequences

The species the reads belonged to was *Podarcis lilfordi*

This are the results of some of the BLAST searches I used to find possible species:

Jan Izquierdo

Edit Search

Save Search

Search Summary

How to read this report?

BLAST Help Videos

Back to Traditional Results Page

Job Title

20 sequences (A00500:270:H7YGVDSX2:1:1101:3224:1000...

RID

UUP1S42G013 Search expires on 01-23 00:45 am Download All

Results for

1:1cl|Query_1974946 A00500:270:H7YGVDSX2:1:1101:3224:1000 1:N:▼

Program

BLASTN Citation

Database

nt See details

Query ID

1cl|Query_1974946

Description

A00500:270:H7YGVDSX2:1:1101:3224:1000 1:N:0:TAAGTATG

Molecule type

dna

Query Length

151

Other reports

Distance tree of results MSA viewer

Filter Results

Organism

only top 20 will appear

☐ exclude

Type common name, binomial, taxid or group name

Add organism

Percent Identity

E value

Query Coverage

to to to

Filter

Reset

Descriptions

Graphic Summary

Alignments

Taxonomy

Sequences producing significant alignments									
Download Select columns Show 100									
select all 100 sequences selected									
GenBank Graphics Distance tree of results MSA Viewer									
	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
<input checked="" type="checkbox"/>	Limanda limanda genome assembly_chromosome_4	Limanda limanda	77.0	197	37%	1e-09	100.00%	32724059	OY754995.1
<input checked="" type="checkbox"/>	Lateolabrax maculatus chromosome Lm14	Lateolabrax mac...	76.1	76.1	32%	4e-09	93.88%	20194087	CP027275.1
<input checked="" type="checkbox"/>	Sardina pilchardus genome assembly_chromosome_6	Sardina pilchardus	76.1	190	54%	4e-09	86.36%	36570839	OY974091.1

BLAST » » blastn suite » results for RID-UU1JGJ09013

Home Recent Results Saved Strategies Help

Edit Search

Save Search

Search Summary

How to read this report?

BLAST Help Videos

Back to Traditional Results Page

Job Title

A00500:270:H7YGVDSX2:1:1101:26151:1000 1:N:0:TAAGTATG

RID

UU1JGJ09013 Search expires on 01-22 18:56 pm Download All

Program

BLASTN Citation

Database

nt See details

Query ID

1cl|Query_3168019

Description

A00500:270:H7YGVDSX2:1:1101:26151:1000 1:N:0:TAAGTATG

Molecule type

dna

Query Length

151

Other reports

Distance tree of results MSA viewer

Filter Results

Organism

only top 20 will appear

☐ exclude

Type common name, binomial, taxid or group name

Add organism

Percent Identity

E value

Query Coverage

to to to

Filter

Reset

Descriptions

Graphic Summary

Alignments

Taxonomy

Sequences producing significant alignments									
Download Select columns Show 100									
select all 42 sequences selected									
GenBank Graphics Distance tree of results MSA Viewer									
	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
<input checked="" type="checkbox"/>	Podarcis lilfordi genome assembly_chromosome_10	Podarcis lilfordi	229	6195	86%	3e-55	99.24%	71836976	OX395135.1
<input checked="" type="checkbox"/>	Podarcis lilfordi genome assembly_chromosome_9	Podarcis lilfordi	229	7424	86%	3e-55	99.24%	77034498	OX395134.1
<input checked="" type="checkbox"/>	Podarcis lilfordi genome assembly_chromosome_8	Podarcis lilfordi	229	6768	86%	3e-55	99.24%	84956682	OX395133.1
<input checked="" type="checkbox"/>	Podarcis lilfordi genome assembly_chromosome_7	Podarcis lilfordi	229	9291	86%	3e-55	99.24%	89641981	OX395132.1
<input checked="" type="checkbox"/>	Podarcis lilfordi genome assembly_chromosome_11	Podarcis lilfordi	229	4456	86%	3e-55	99.24%	62796525	OX395136.1
<input checked="" type="checkbox"/>	Podarcis lilfordi genome assembly_chromosome_16	Podarcis lilfordi	229	1556	86%	3e-55	99.24%	41280576	OX395143.1
<input checked="" type="checkbox"/>	Podarcis lilfordi genome assembly_chromosome_15	Podarcis lilfordi	229	3622	86%	3e-55	99.24%	44568424	OX395141.1
<input checked="" type="checkbox"/>	Podarcis lilfordi genome assembly_chromosome_W	Podarcis lilfordi	229	26267	90%	3e-55	99.24%	12316893	OX395145.1
<input checked="" type="checkbox"/>	Podarcis lilfordi genome assembly_chromosome_14	Podarcis lilfordi	224	2824	86%	3e-54	98.47%	53023510	OX395139.1
<input checked="" type="checkbox"/>	Podarcis lilfordi genome assembly_chromosome_13	Podarcis lilfordi	224	3652	86%	3e-54	98.47%	53850232	OX395138.1
<input checked="" type="checkbox"/>	Podarcis lilfordi genome assembly_chromosome_17	Podarcis lilfordi	224	2337	86%	3e-54	98.47%	41611985	OX395142.1
<input checked="" type="checkbox"/>	Podarcis cretensis genome assembly_chromosome_14	Podarcis cretensis	224	2685	86%	3e-54	98.47%	56358416	OX638162.1
<input checked="" type="checkbox"/>	Podarcis cretensis genome assembly_chromosome_13	Podarcis cretensis	224	2493	86%	3e-54	98.47%	57670597	OX638161.1