Martí Díez Macià
NIA: **107722**

 # Ex 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)

```
fastq_path = '/home/marti/Baixades/unknown_illumina_2024.fastq'
f = open("/home/marti/Baixades/q8.fasta", "w")
        for read in SeqIO.parse(fastq_path, "fastq"):
                print(">"+str(read.id), file=f)
                print(read.seq, file=f)
```

Step by step code function:

1. Read a working directory and assign it to a variable named fastq_path
2. Open a file called q8.fasta in order to write on it and assign it to a variable named f
3. For each line in the fastq file, write the header and then the sequence from the file into the new file called q8.fasta

Once I got the q8.fasta done, I went to the NCBI BLAST website in order to find where the reads originated from. The BLAST tool (Basic Local Alignment Search Tool) finds regions of local similarity between sequences. It is an algorithm used for comparing biological sequence information, in our case the nucleotide sequences of DNA. It is used to help understand the evolutionary relationships between them, among other uses. I used BLASTN to compare a nucleotide query sequence against a nucleotide sequence database.

The first read suggested that the genome was from a fish called Limanda limanda, on the 4th chromosome as we can interpret from these results

☑	Limanda limanda genome assembly, chromosome: 4	Limanda…	77.0	197	37%	1e-09	100.00%

But with further research I found out that it came from a lizard called Podarcis lilfordi, on the 9th chromosome

☑	Podarcis lilfordi genome assembly, chromosome: 9	Podarcis…	221	221	84%	4e-53	98.43%

I reached this conclusion since the query cover is higher on the Podarcis lilfordi (84% against 37%) and also since the e-value is far smaller on the Podarcis lilfordi (4e-53 < 1e-09). A high identity percentage and a low e-value clearly indicates that the sequences are related, and in our case it shows that it belongs to the Podarcis lilfordi species, also known as Lilford's wall lizard.

What this data shows is that it is more plausible that our sequence belongs to the Podarcis lilfordi species, and not to the Limanda limanda species.

The methods used in order to find the species have been the following:
Python: to change the file format from .fastq to .fasta
Nucleotide BLAST search: to compare the sequence obtained from the .fasta file to the databases on the NCBI website