# CINEMA Paris 2023

# Environmental microbiome analysis hands-on demo Part 1

Form small groups of 3-4. Introduce yourselves, your background research interests, and experience.

You have an assembled metagenome from a wastewater bioreactor. For your research you want to assess the antibiotic resistance profile within this sample.

Input files:

- Assembled nucleotide sequences generated from the sample "nucl.contigs.fa"
- ARG database, megares_database_v3.00.fasta
  *Information on the database is available here. https://www.meglab.org/megares/*
  *In theory you could apply this workflow to any databases of interest. Or you could make a custom database of sequences of interest to your research questions.*

Install standalone BLAST

On a Mac machine download BLAST (.macosx.tar.gz file) here:
https://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/
Instructions for installation on a PC are available here:
https://www.ncbi.nlm.nih.gov/books/NBK52637/

Move the folder associated with BLAST to your working folder where the files are. Open a terminal window to run commands and navigate to your working directory (where the files and BLAST are downloaded).

First format the database fasta file so that it can be used as a database.

```
ncbi-blast-2.14.0+/bin/makeblastdb -in megares_database_v3.00.fasta -dbtype nucl -out
                ..
```

Output should look something like this:

```
Building a new DB, current time: 06/29/2023 17:36:27
New DB name:   /Users/jesethdv/Delgado Vela lab Dropbox/Jeseth Delgado Vela/Teaching/CINEMA/Supplemental Materials/megares_db
New DB title:  megares_database_v3.00.fasta
Sequence type: Nucleotide
Keep MBits: T
Maximum file size: 3000000000B
Adding sequences from FASTA; added 8733 sequences in 0.113232 seconds.


BLAST Database error: No alias or index file found for nucleotide database [/Users/jesethdv/Delgado] in search path [/Users/jesethdv/Delgado Vela lab Dropbox/Jeseth Delgado Vel
a/Teaching/CINEMA/Supplemental Materials::]
```

Compare the database against the nucleotide sequences in the assembly using this command:

```
ncbi-blast-2.14.0+/bin/blastn -query nucl.contigs.fa -db megares_db -outfmt "6 std qlen" -out
ARG.blasta
```

- Decide what types of 'hits' are acceptable. LINK TO RESOUCES

**Point of discussion in your small groups**
An alternative analysis would be to turn your nucleotide sequences in your assembly to protein coding sequences using tools such as prodigal, then run tblastn to compare the translated sequences to your nucleotide database.
- What are some advantages of disadvantages of each approach?

You can import the ARG.blastn to your preferred viewer for tab delimited files (e.g., R, a Python IDE, or Excel). The columns of the file are:

assembly identifier; database identifier; percent ID; alignment length; mismatch; gap opening; query start; query end; subject start; subject end;
e-value; bit score; query length

Filter your BLAST outputs to eliminate redundancies and include only acceptable hits. Save this as a new file. As a group, develop cutoffs for quality of hit (e-value and bit scores or length).

Next, you have a few MAGs from this sample in the folder called MAGs. The bins were generated using CONCOCT, which uses both tetranucleotide frequency and coverage to generate bins. The folder also includes outputs from dRep from this study.

Open the outputs and familiarize yourself with the information in the outputs. Find the MAGs that you were given within the output.

**Point of discussion in your small groups**
What level of completeness/contamination is acceptable for your research questions? How would you interpret the 'quality' of the MAGs you have been given?

# Environmental microbiome analysis hands-on demo Part 2

In your same groups, let's put it together and link MAGs with ARG hits.

One way to do this would be to search the ARG hits file you developed and compare to the MAGs. This is where you can practice the majority of what you will be doing with metagenomic analysis- data wrangling and comparing disparate files!

My (very inelegant) solution in command line was to take my ARG hits and make a new file with just the column of the contigs that had hits, then compare that to each bin. Of course, if you have many bins you would write code to go through all of your bins. You should do this on whatever your filtered file your group generated with 'good' hits to the database.

```
cut -ft -d$'\t' ARG.blastn > hits.txt

grep -f hits.txt 12_5_biomass_bin_2.fa > bin_2_ARGContigs.txt

grep -f hits.txt 12_5_biomass_bin_15.fa > bin_15_ARGContigs.txt
```

If you are more familiar or comfortable with R that can be somewhere else you do this sort of analysis.

**Points of discussion in your small groups**
Of course, things aren't quite this easy. Look through some of these entries in the database, what are some limitations that you notice (HINT: 23S and rpoB…hmmmmmmm)?

How would you decide if these MAGs have any clinical relevance?

If time allows, find a bin from an ecosystem of interest to you on JGI. Explore JGI as a tool for downloading genomes that you could explore further.
https://img.jgi.doe.gov/cgi-bin/m/main.cgi?section=MetagenomeBins&page=bins&type=ecosystem