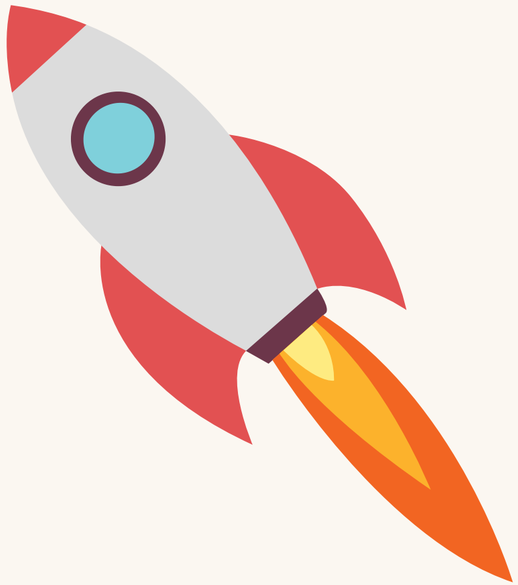
Reference Database & BLAST



Where are we now?

- We have concensus sequences representing ‘unique’ groups of reads, AKA MOTUs
- The next step is to match these MOTUs to reference sequenes. The way we will do this matching is through BLAST.
- To match to a reference sequences (BLAST), we need to first create a reference database.

Reference Database

Reference Database

Defined as a set of reference sequences (i.e., sequences where we know what species it is).

Different types:

1. Global reference db - all known reference sequences
2. Local reference db - a subset of all known reference sequences filtered to species only found in survey area.
3. Regional reference db - same as local except extended to a broader region.

What to do?

- First compare MOTUS vs. local
- Then compare any unmatched MOTUs to global.
- Why?

How to Create a Local Reference Database?

- It can be a detailed and technical process
- I have some resources for using `crabs` - a widely used workflow to create a local reference database of amplicons.
- However, for this workshop we will do it a simpler way*
  - Use an existing curated ‘global’ 12S reference database called MIDORI2
  - Subset it to species from our survey area (fish found around Naples, Florida)

*which serves as a gentle introduction to reference db creation via a method that is not quiet suitable for publication.

Simplified Local Reference Database Creation Steps

1. Generate list of local species (GBIF)
2. Download MIDORI2, 12S fasta
3. Filter MIDORI to local species
4. Convert above fasta to a blast database format (needed for BLAST)

Generate list of local species (GBIF)

1. Generate a list of species found in study area. This can be done using GBIF or any other data source of just from expert knowledge. Note GBIF has an API, but it's simple enough just to use the website.
    - To use GBIF, first go to https://www.gbif.org/
    - First, login with your username (or else you'll need to repeat all the below steps twice).
    - Go to Occurrences
    - Under 'Search all fields', type in the taxa of interest; eg Amphibia and press the Enter key (make sure it's the right taxonomic rank, eg Class). I recommend typing in 'Chordata' (especially when you need fish), it should prompt you with this "Your search matches a Phylum: 'Chordata'. Do you wish to limit your search to this taxon only?", click Yes.
    - Under Occurrence Status, make sure "Present" occurrences are selected and "Absent" are not selected.
    - Subset by location. You can subset occurrences by locations by using one of the following, a GIS file (e.g. shapefile), a freehand drawn polygon/square on a map, or by Name (e.g., state name).
        - To use a GIS file, it must be in geojson format. If it's not currently, you can use an online converter. Under Location, click Geometry and paste the contents of the geojson file there and click ADD. If the file is "too large" it will give you a message and you can simplify it. Make sure the shape still looks correct.
        - To subset by name, I recommend using the "Administrative areas (gadm.org) filter. You can type in the name of interest (e.g., New Jersey)
    - You can browse the other filters to see if any apply to you. For instance, you might want to only have observations after a certain date/year. You can browse the distribution of observation by year under the "METRICS" panel.
    - To download the species list, go to the "DOWNLOAD" panel. Then click download "SPECIES LIST". The action should show "Under Processing". The species list will NOT be immediately available, but GBIF will email you when ready (usually 1-3 hours).

https://www.gbif.org/

I've done this already for fish near Naples, Florida

Download MIDORI2, 12S fasta

https://www.reference-midori.info

Download:

```
wget https://www.reference-midori.info/download/Databases/GenBank261_2024-06-15/RAW/uniq/MIDORI2_UNIQ_NUC_GB261_srRNA_RAW.fasta.gz \
    -O data/midori/midori-12S.tar.gz
```

Unzip:

```
gzip -d -c $crabs_dl_dir/midori/midori-12S.tar.gz > \
    data/midori/midori-12S.fasta
```

Filter MIDORI to local species

To do this, we will use a function from crabs.

- First, convert the MIDORI fasta to crabs format (tsv)
- Run crabs filter which takes input of file of local species
- Convert to .fasta and .csv to be our local reference database

Convert to a blast database format

A blast database is the format that BLAST uses (not fasta)

```
singularity exec images/blast.sif \
  makeblastdb -in "local_refdb.fasta" \
  -parse_seqids \
  -taxid_map "taxid_map.txt" \
  -dbtype nucl \
  -out "blastdb/name_of_blastdb/name_of_blastdb"
```

BLAST is very tricky to set up properly. In order to get taxonomy in the output of BLAST, we need also do the following:

- Generate a taxid map (before running makeblastdb
- Download NCBI taxonomy files and add to blastdb dir (after running makeblasdb)

BLAST

BLAST

BLASTn

- compares one or more nucleotide query sequences to a database of nucleotide sequences.
- Good for checking species identities based on reference gene sequence(s)
- accounts for small indels

Anaxyrus woodhousii voucher UTA 53926 tRNA-Phe gene, partial sequence; 12S ribosomal RNA and tRNA-Val genes, complete sequence; and 16S ribosomal RNA gene, partial sequence; mitochondrial

Sequence ID: AY680216.1 Length: 2430 Number of Matches: 1

Range 1: 855 to 908 GenBank Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Gaps	Strand
95.3 bits(51)	4e-16	54/55(98%)	1/55(1%)	Plus/Plus

Consensus sequence from metabarcoding run -->

Reference Sequence -->

```
Query  1    CTTCAAAGCTAATCTAATCTAGTTCTTAACATATTTAAGACCTTACAGAAGAGGC  55
            ||||||||||||||||||||||||||||||||||||||| ||||||||||||||||
Sbjct  855  CTTCAAAGCTAATCTAATCTAGTTCTTAACATATTTAAG-CCTTACAGAAGAGGC  908
```

BLAST Output Columns

BLASTn tabular output format 6

Column headers:

```
qseqid sseqid pident length mismatch gapopen qstart qend sstart send evalue bitscore
```

1. **qseqid** query or source (gene) sequence id
2. **sseqid** subject or target (reference genome) sequence id
3. **pident** percentage of identical positions
4. **length** alignment length (sequence overlap)
5. **mismatch** number of mismatches
6. **gapopen** number of gap openings
7. **qstart** start of alignment in query
8. **qend** end of alignment in query
9. **sstart** start of alignment in subject
10. **send** end of alignment in subject
11. **evalue** expect value
12. **bitscore** bit score

https://www.metagenomics.wiki/tools/blast/blastn-output-format-6

Next Steps?

- Clean up BLAST output
  - High pident (species level, genus level)
  - Min aligment length
  - Add back in number of reads (remember we blasted consensus seqs)
- Account for contaminants, negatives, and postive controls
  - Hint, we can’t do that with this dataset