MAKING A

LOCAL REFERENCE DATABASE

THE LEGIT WAY*

*Disclaimer: Oliver is just learning this and doesn't know everything. The same can be said about everything Oliver says, but even more so here.

What is a local reference database?

1. A set of reference sequences (i.e., sequences where we know what species it is) …
2. at the amplicon of interest of …
3. the species/taxa ‘you’ know are present at the sampling site

WHY CREATE A LOCAL REF DB?

What's wrong with just comparing against all known reference sequences?

- For metabarcoding, it's the best way to get the most accurate species identifications.
- By restricting to only species present in the sampling site, you can narrow down matches of DNA with less taxa compared to a global match.
- Output is a realistic set of identifications based on prior knoledge of the site.
- (Metabarcoding should probably **not** be used to ID new species to a location, even though it's tempting).

COMMON SCENARIO

Not all species in sample site have a reference sequence for a given amplicon.

- This is a reality for a lot of taxa and locations, although US has good coverage for a lot of taxa/regions.
- Species A with no reference sequence might match up to Species B (in same Genus) at 99%, but we would never know because we would just call it Species B.
- This is a limitation of the metabarcoding, primer choice, and of data availability of reference sequences.
- If this is a huge problem, you need to think of other solutions like using different regions/primers.
  - which implies making a local refence database should be done before choosing a primer … ?
- If ‘unknown’ species is different enough to sister taxa, can just include all genus sequences and if there is a hit, we can conclude it’s within the genus.

STEPS

1. Generate a list of species found in study area
2. Verify species in local species list with an expert to either add or remove species
3. Align the taxonomy of the local species list to NCBI taxonomy
4. Curate custom sequences
5. Make sure to include positive control species and 'contaminant' species in the above steps.

For entire project

6. Download reference sequences

For a gene (ish)

7. Run crabs workflow (primer specific)
    a. in silico PCR on reference sequences given primer set used
    b. pairwise global alignment (PGA) of unmatched sequences
    c. crabs filter: to local species list + other filtering steps
8. More curation and quality control

For a primer

LIST OF LOCAL SPECIES

1. **Generate a list of species found in study area**
2. **Verify species in local species list with an expert to either add or remove species**

- This determines the possibility of what references sequences we gather and ultimately what species can be considered a match.
- I've shown one method for #1 via GBIF but there are other data sources.
- #2 is a seperate step because it's important. I think we should be working with experts and/or the client on this step together.
  - If they're looped in from the beginning, you will have the most solid list before analysis. Then when you receive your 'final' species list, there is no questioning if a species belongs.
  - Basically, in every metabarcoding project the question of whether a species is found at a location is always going to be a question you need to answer, so I argue it should be tackled early and systematically.
  - It's never that easy, of course, and it can be a more iterative process by seeing the results and adjusting as needed. But giving it a a good attempt will go a long way.

3. ALIGN THE TAXONOMY OF THE LOCAL SPECIES LIST TO NCBI TAXONOMY

Why?

- Sometimes species names in other databases are different than the species name that NCBI uses
- Since NCBI is the main source for the reference seqs, the species names need to match NCBI names.
  - If the species name does not match, it will not end up in our local reference database, despite there being reference sequences.
- If EMBL is the main source, should also align with EMBL names, etc.

How?

- Use R to automate,
  - I have a script. It loops over each species name and looks it up to see if it matches NCBI, can also check for synonyms (i think)
- OR manually look up

4. Curate custom sequences

Sequences you have that are not (yet) published can be included in your local reference database.

How?

- First, make a csv of them, including: a custom accession number, NCBI species name, NCBI taxid, and the sequence
- Later we will append this to our output.

seqID	species	taxid	sequence
RU_12S_Pmontanus_01	Pseudotriton montanus	324349	TGCCAGACA(
RU_12S_Tuna_01	Thunnus	8234	GTGCCAGCC
RU_12S_Cstriata_01	Centropristis striata	184440	TTTTCTGTTG(
RU_12S_Ochrysurus_01	Ocyurus chrysurus	40499	GCCAGCCAC

5. MAKE SURE TO INCLUDE POSITIVE CONTROL SPECIES AND 'CONTAMINANT' SPECIES IN THE ABOVE STEPS.

These need to be in the local reference database. Positive Control for tag jumping and Contaminants for ?

How?

1. Make 2 .txt files for each where each line is a species -->
2. Later on, we'll feed these species names into the workflow to make the reference db.

contaminant_species.txt

```
1  Homo sapiens
2  Pan troglodytes
3  Gallus gallus
4  Felis catus
5  Canis lupus
6  Rattus rattus
7  Mus musculus
8  Equus asinus
9  Equus caballus
10 Ovis aries
11 Sus scrofa
12 Bos taurus
13 Capra hircus
```

6. DOWNLOAD REFERENCE SEQUENCES

crabs has easy functions for this.

For NCBI, we need to make a query:

```
### birds
crabs --download-ncbi \
--query \
'("Aves"[Organism] OR Aves[All Fields]) AND \
AND ("Tetrapoda"[Organism] OR Tetrapoda[All Fields]) \
AND ("mitochondrion"[filter] OR "mitochondrial genome"[All Fields]) \
AND ("80"[SLEN] : "20000"[SLEN])' \
--output $crabs_dl_dir/birds/ncbi_birds_dl_$date.fasta \
--email $email \
--database nucleotide
```

For MIDORI2, we need to know the gene of interest (12S, 16S, COI)

Other databases (Mitofish, etc.) might have other specifications.

7. RUN CRABS WORKFLOW

crabs has easy functions for this.
This is done seperately for each primer.

1. **In silico PCR of downloaded sequences** (crabs uses cutadapt behind the scenes)
   - Note, I also do 2 extra in silico PCRs for 'unlinked' sequences (only one primer end match)

```
crabs --in-silico-pcr
--input testing/merged.txt
--output testing/insilico.txt
--forward GACCCTATGGAGCTTTA
--reverse CGCTGTTATCCCTADRGTAA
```

2. **pairwise global alignment of remaining sequences** (crabs uses vsearch behind the scenes)

```
crabs --pairwise-global-alignment
--input testing/merged.txt
--amplicons testing/insilico.txt
--output testing/aligned.txt
--forward GACCCTATGGAGCTTTA
--reverse CGCTGTTATCCCTADRGTAA
--size-select 10000
--percent-identity 0.95
--coverage 95
```

3. **crabs filtering**
   - *subset to local species + contams + positive control*
   - min/max length; number of Ns; if ref seq is 'environmental'; remove seqs w/no species id; remove seqs with >1 NA in taxonomy

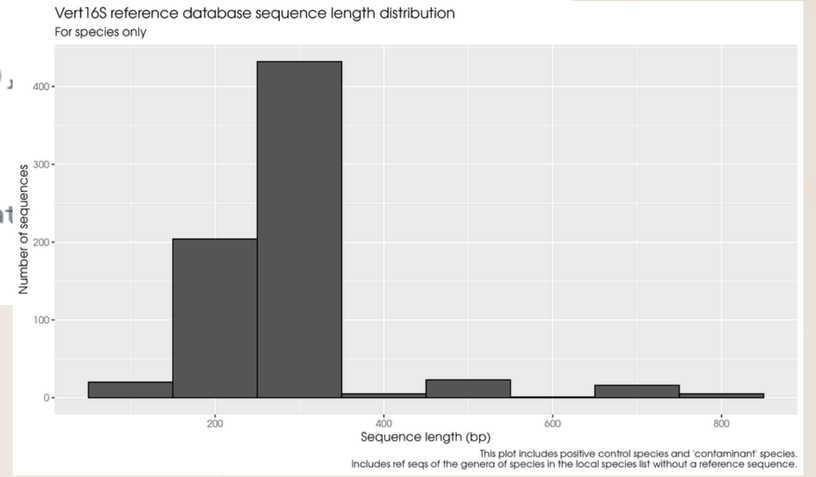
https://github.com/gjeunen/reference_database_creator

8. More curation and quality control

Automated in a script:

```
# REQUIRES INTERNET to look up genus taxid

# This script does the following:
# 1. Adds in custom seqs
# 2. Checks if accession numbers belongs to the correct gene or whole mitogenome
# 3. Checks if any local species have no reference sequences
# 4. If there are no reference sequences for a given local species,
##   the script includes reference seqs of sister taxa (if available) and calls it a genus level reference sequence, using the genus taxid
##   Why do this? This will help match reads of the missing species to the genus level ref seq.
# 5. Removes duplicates, defined as records with the same accession number and is a 100% subset of another record with the same accession number
##   The order of priority to keep these records is as follows: insilico seqs (from crabs) followed by the longest seq. If duplicated
##   If duplicated accessions do not have an insilico match, and one is not a 100% overlap (ie different region on gene),
##   the one with the shortest length is kept.
# 6. Removes duplicated sequences within a species (ignoring acession numbers)
# 7. For contaminant species, only keep sequences with at least 10 exact duplicates (from above step),
##   for humans at least 100 duplicates. Note, if the contaminant species does not have at least 10 (unlikely),
##   then all sequences are kept. This is done because there are so many sequences for contaminant species,
##   and the ones that are not duplicates are more likely to be lower quality / mis labelled in genbank.
# 8. Makes a final refdb for local species, positive control and contaminants - based on all of the above curat
# 9. Generates summary files and figures
```



Then finally export as a .fasta and .csv to be used in the bioinformatics pipeline

END

NEXT WE WILL CODE IT