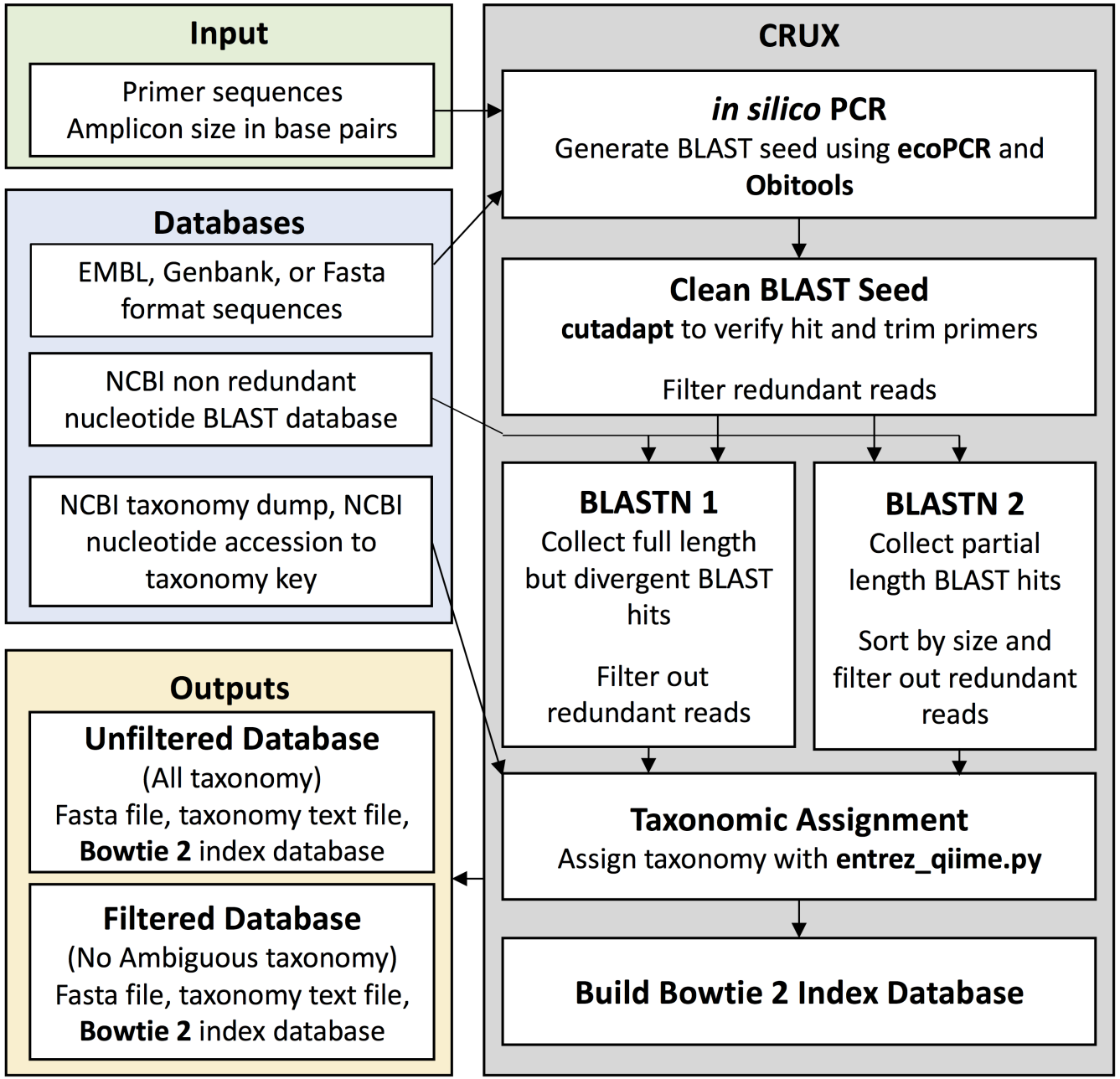
CRUX: Creating Reference libraries Using eXisting tools

Workshop 10/16/2020

Lenore Pipes (UC-Berkeley Postdoc – Nielsen lab)

CRUX: **C**reating **R**eference libraries **U**sing e**X**isting Tools



- ecoPCR forward_primer reverse_primer from EMBL libraries
- Remove primer sequences
- 2 blast steps for each BLAST seed
- Converts accession to taxonomic paths
- bowtie2-build builds the bowtie2 database which is used in Anacapa QC for taxonomic assignment

What you need installed for the workshop

- `obiconvert`
  `obiconvert -h`
- `ecoPCR`
  `ecoPCR –h`
- `cutadapt`
  `cutadapt –help`
- `blastn` (make sure version is compatible with NCBI nt database)
  `blastn –h`
- `bowtie2-build`
  `bowtie2-build`

Databases needed for CRUX

- NCBI taxonomy dump

```
wget ftp://ftp.ncbi.nlm.nih.gov/pub/taxonomy/taxdump.tar.gz
mkdir TAXO
mv taxdump.tar.gz
cd TAXO
tar xvzf taxdump.tar.gz
```

Databases needed for CRUX

- NCBI accession2taxonomy file

  wget ftp://ftp.ncbi.nih.gov/pub/taxonomy/accession2taxid/nucl_gb.accession2taxid.gz

  ```
  gunzip nucl_gb.accession2taxid.gz
  ```

Databases needed for CRUX

- NCBI nt database

```
mkdir NCBI_blast_nt
cd NCBI_blast_nt
wget ftp://ftp.ncbi.nlm.nih.gov/blast/db/nt*
for file in nt*.tar.gz; do tar -zxf $file; done
```

Downloading the EMBL libraries (used for ecoPCR as the “BLAST seed”)

If you go to ftp://ftp.ebi.ac.uk/pub/databases/embl/release/std/ in your web browser, you can view all of the std libraries:

Division	Code
Bacteriophage	PHG - common
Environmental Sample	ENV - common
Fungal	FUN - map to PLN (plants + fungal)
Human	HUM - map to PRI (primates)
Invertebrate	INV - common
Other Mammal	MAM - common
Other Vertebrate	VRT - common
Mus musculus	MUS - map to ROD (rodent)
Plant	PLN - common
Prokaryote	PRO - map to BCT (poor name)
Other Rodent	ROD - common
Synthetic	SYN - common
Transgenic	TGN - ??? map to SYN ???
Unclassified	UNC - map to UNK
Viral	VRL - common

- std library is only a PARTIAL sequence library and may not capture your target species
- it is also possible to add the wgs library (whole genome shotgun) to include more sequences for your BLAST seed

Downloading the EMBL libraries (used for ecoPCR as the “BLAST seed”)

```
mkdir Obitools_databases

cd Obitools_databases

wget ftp://ftp.ebi.ac.uk/pub/databases/embl/release/std/rel_est_mam_*1*_r143.dat.gz
```

Use this for-loop to create directories for each EMBL *.dat file, gunzip, obiconvert, and then remove the *.dat file

```
for file in rel_est_mam_*.gz; do name=`basename
$file .dat.gz`; mkdir OB_dat_EMBL_${name};
gunzip $file; obiconvert -t
/space/s1/lenore/TAXO --embl --ecopcrdb-
output=OB_dat_EMBL_${name}/OB_dat_EMBL_${name}
${name}.dat; rm ${name}.dat; done
```

- First download CRUX (if you haven't done so already)

```
git clone https://github.com/limey-bean/CRUX_Creating-Reference-libraries-Using-eXisting-tools.git
```

```
wget https://github.com/limey-bean/CRUX_Creating-Reference-libraries-Using-eXisting-tools/archive/master.zip
```

- Change the directory

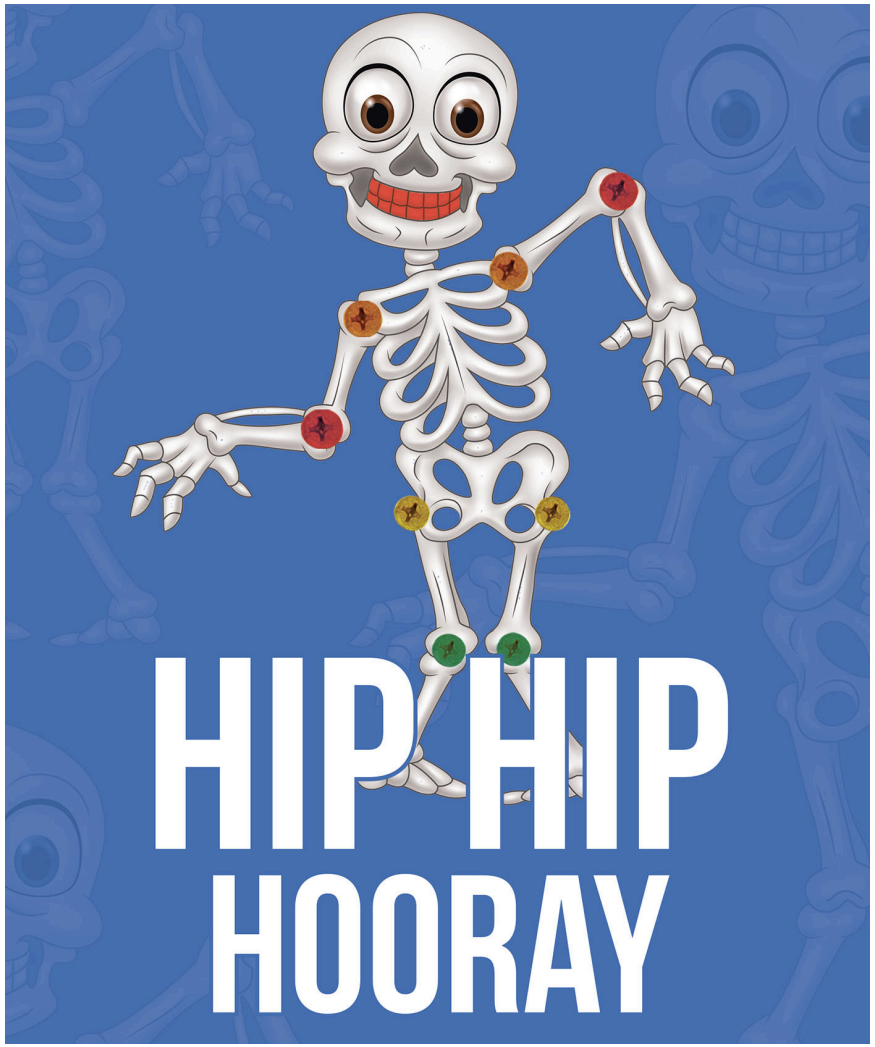
```
cd CRUX_Creating-Reference-libraries-Using-eXisting-tools
```

Modifying the crux_config.sh file

- Change MODULE_SOURCE=“”
- Change CUTADAPT=“cutadapt”
- Change FASTX_TOOLKIT=“” (doesn’t actually use this)
- Change ecoPCR=“ecoPCR”
- Change BLASTn_CMD=“blastn”
- Change QIIME=“”
- Change BOWTIE2=“”
- Change ATS=“”
- Change OBI_DB=“/space/s1/lenore/crux_workshop/Obitools_databases”
- Change TAXO=“/space/s1/lenore/TAXO”
- Change A2T=“/space/s1/lenore/ncbi-nt/nucl_gb.accession2taxid”
- Change BLAST_DB=“/space/s1/lenore/ncbi-nt/nt”

OK! Now we are ready to run CRUX

```
bash crux.sh -h
```



You can also change default options by changing the crux_vars.sh file

```
<<< CRUX: Sequence Creating Reference Libraries Using eXisting tools>>>

The purpose of these script is to generate metabarcode locus specific reference libraries. This script takes P
        For successful implementation
                1. Make sure you have all of the dependencies and correct paths in the crux_config.sh file
                2. All parameters can be modified using the arguments below.  Alternatively, all parameters ca

Arguments:
- Required:
        -n      Metabarcode locus primer set name
        -f      Metabarcode locus forward primer sequence
        -r      Metabarcode locus reverse primer sequence
        -s      Shortest amplicon expected (e.g. 100 bp shorter than the average amplicon length
        -m      Longest amplicon expected (e.g. 100 bp longer than the average amplicon length
        -o      path to output directory
        -d      path to crux_db

- Optional:
        -x      If retaining intermediate files: -x (no argument needed; Default is to delete intermediate fil
        -u      If running on an HPC (e.g. UCLA's Hoffman2 cluster), this is your username: e.g. eecurd
        -l      If running locally: -l  (no argument needed)
        -k      Chunk size for breaking up blast seeds (default 500)
        -e      Maximum number of mismatch between primers and EMBL database sequences (default 3)
        -g      Maximum number of allowed errors for filtering and trimming the BLAST seed sequences with cuta
        -t      The number of threads to launch for the first round of BLAST (default 10)
        -v      The minimum accepted value for BLAST hits in the first round of BLAST (default 0.00001)
        -i      The minimum percent ID for BLAST hits in the first round of BLAST (default 50)
        -c      Minimum percent of length of a query that a BLAST hit must cover (default 100)
        -a      Maximum number of BLAST hits to return for each query (default 10000)
        -z      BLAST gap opening penalty
        -y      BLAST gap extension penalty
        -j      The number of threads to launch for the first round of BLAST (default 10)
        -w      The minimum accepted value for BLAST hits in the first round of BLAST (default 0.00001)
        -p      The minimum percent ID for BLAST hits in the first round of BLAST (default 70)
        -q      Minimum percent of length of a query that a BLAST hit must cover (default 70)
        -b      HPC mode header template
```

Toy example using CO1 Leray et al. primers

```
mkdir crux_output


bash CRUX_Creating-Reference-libraries-Using-
eXisting-tools/crux_db/crux.sh -l -n CO1 -f
GGWACWGGWTGAACWGTWTAYCCYCC -r
TANACYTCnGGRTGNCCRAARAAYCA -s 0 -m 2000 -o
crux_output -d CRUX_Creating-Reference-
libraries-Using-eXisting-tools/crux_db
```

Downloading the EMBL libraries (used for ecoPCR as the “BLAST seed”)

Change

wget ftp://ftp.ebi.ac.uk/pub/databases/embl/release/std/rel_est_mam_*1*_r143.dat.gz

To

ftp://ftp.ebi.ac.uk/pub/databases/embl/release/std/*.dat.gz

Use this for-loop to create directories for each EMBL *.dat file, gunzip, obiconvert, and then remove the *.dat file


```
for file in rel_est_mam_*.gz; do name=`basename
$file .dat.gz`; mkdir OB_dat_EMBL_${name};
gunzip $file; obiconvert -t
/space/s1/lenore/TAXO --embl --ecopcrdb-
output=OB_dat_EMBL_${name}/OB_dat_EMBL_${name}
${name}.dat; rm ${name}.dat; done
```